2020

Statistical Methods For Multi-Omics Inference From Single Cell Transcriptome

Zilu Zhou
University of Pennsylvania

Statistical Methods For Multi-Omics Inference From Single Cell Transcriptome

Abstract

This thesis comprises three sections of research in statistical genomics and computational biology. Chapter 1 and Chapter 2 describe two statistical methods for multi-omics inference from single cell transcriptome, representing the theme of this thesis. Chapter 3 describes a side-project on copy

number variation detection in large biobank data base.

Part 1: Although scRNA-seq is now ubiquitously adopted in studies of intratumor heterogeneity, detection of somatic mutations and inference of clonal membership from scRNA-seq is currently unreliable. We propose DENDRO, an analysis method for scRNA-seq data that detects genetically distinct subclones, assigns each single cell to a subclone, and reconstructs the phylogenetic tree describing the tumor's evolutionary history. DENDRO utilizes information from single nucleotide mutations in transcribed regions and accounts for technical noise and expression stochasticity at the single cell level. The accuracy of DENDRO was benchmarked on spike-in datasets and on scRNA-seq data with known subpopulation structure. We applied DENDRO to delineate subclonal expansion in a mouse melanoma model in response to immunotherapy, highlighting the role of neoantigens in treatment response. We also applied DENDRO to primary and lymph-node metastasis samples in breast cancer, where the new approach allowed us to better understand the relationship between genetic and transcriptomic intratumor variation.

Part 2: Recent technological advances allow the simultaneous profiling, across many cells in parallel, of multiple omics features in the same cell. In particular, high throughput quantification of the transcriptome and a selected panel of cell surface proteins in the same cell is now feasible through the REAP-seq and CITE-seq protocols. Yet, due to technological barriers and cost considerations, most single cell studies, including Human Cell Atlas (HCA) project, quantify the transcriptome only and do not have cell-matched measurements of relevant surface proteins that can serve as integral markers of cellular function and targets for therapeutic intervention. Here we propose cTP-net (single cell Transcriptome to Protein prediction with deep neural network), a transfer learning approach based on deep neural networks, that imputes surface protein abundances for scRNA-seq data. Through comprehensive benchmark evaluations and applications to HCA and AML data sets, we show that cTP-net outperform existing methods and can transfer information from training data to accurately impute 24 immunophenotype markers, which achieve a more detailed characterization of cellular state and cellular phenotypes than transcriptome measurements alone. cTP-net relies, for model training, on accumulating public data of cells with paired transcriptome and surface protein measurements.

Part 3: Copy number variations (CNVs) are gains and losses of DNA segments that are highly associated with multiple diseases. The Penn Medicine BioBank stores SNP-array and NGS data for more than 10000 individuals across ethnicity and conditions, providing a rich resource for CNV discovery and analysis. This type of experiment design fits perfectly for CNV detection tool - Integrated Copy Number Variation caller (iCNV), which I developed as my master thesis. The distinguishing feature of iCNV includes adaptation of platform specific normalization, utilization of allele specific reads from sequencing and integration of matched NGS and SNP-array data by a Hidden Markov Model (HMM). We applied iCNV on Penn Medicine BioBank data set, calling CNV over more than 10000 individuals (~2000 AFR, ~8000 EUR) with different phenotypes. iCNV detected on average 34.1 deletions and 11.3 duplications per EUR sample, and 38 deletions and 10.6 duplications per AFR sample. iCNV calling results show great improvement in detection sensitivity and specificity comparing to single platform detection method. Penn Medicine BioBank CNV sets by iCNV provide a rich database for researchers to study the relationship between diseases phenotypes and CNV across ethnicity and conditions.

STATISTICAL METHODS FOR MULTI-OMICS INFERENCE FROM SINGLE CELL TRANSCRIPTOME

Zilu Zhou

A DISSERTATION

in

Genomics and Computational Biology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2020

Supervisor of Dissertation

_____

Nancy R. Zhang

Professor of Statistics; Vice Dean of Wharton Doctoral Programs; Ge Li and Ning Zhao Professor

Graduate Group Chairperson

_____

Benjamin F. Voight

Associate Professor of Genetics, Systems Pharmacology and Translational Therapeutics

Dissertation Committee

Chair: Kai Tan, Associate Professor of Pediatrics, Genetics, and Cell and Developmental Biology

Andy J. Minn, Associate Professor of Radiation Oncology

Mingyao Li, Professor of Biostatistics

Pei Wang, Professor of Genetics and Genomic Sciences

STATISTICAL METHODS FOR MULTI-OMICS INFERENCE FROM SINGLE CELL TRANSCRIPTOME

To Chaoxiu, Xiao, Guiying and Nan

ACKNOWLEDGMENT

I would like to first thank my advisor Nancy Zhang. Nancy is not only an incredible researcher, statistician and visionary in the academic field, but also a wonderful mentor, who has taught me many life lessons and helped me become a better human being. I feel very lucky to have her as my advisor and very grateful for all her support and advises in the past five years.

I would extend my sincere thanks to my thesis committee members, Kai Tan, Andy Minn, Mingyao Li and Pei Wang for their insights and suggestions that guide me through my Ph.D dissertation. I am also grateful to my collaborators Yuchao Jiang, Li-San Wang, Weixin Wang, Mingyao Li, Andy Minn, Bihui Xu, Maja Bucan, Dan Rader, Hanlee Ji, Jimmie Ye, Derek Oldridge and John Wherry, who are willing to share their data and provide constructive advice for my thesis and projects.

I thank Genomics and Computational Biology program, especially Li-San Wang and Ben Voight for recruiting me and guiding me through ups and downs, Maureen Kirsch and Hannah Chervitz for keeping GCB running smoothly, and GCB students and friends that lift each other to go through challenging moments of pursing the degree. I am also grateful to all the professors office mate, and students in the Statistics Department, especially Jian Ding and Mark Low for your helps in the past five years.

None of this work will be possible without my family, especially my parents. Thank you Chaoxiu and Xiao, who have always been supportive to my dream. Thank you my grandma Guiying who taught me the importance of education. I also want to thank my friends, Yaozixin group, Penn Chinese Basketball Association, Penn Biotech Consulting Group, Penn Center of Innovation, Penn Data Science Group and Janssen Pharmaceuticals. You have enlightened my time at Penn with colorful activities. At last, Nan, thank you for being my companions. I feel truly lucky to spend the past three years together and hope many more to come.

Pursing Ph.D. was never an easy path. I feel truly lucky to studied at Penn and thankful to everyone who has lifted me a hand.

ABSTRACT

STATISTICAL METHODS FOR MULTI-OMICS INFERENCE FROM SINGLE CELL TRANSCRIPTOME

Zilu Zhou

Nancy R. Zhang

This thesis comprises three sections of research in statistical genomics and computational biology. Chapter 1 and Chapter 2 describe two statistical methods for multi-omics inference from single cell transcriptome, representing the theme of this thesis. Chapter 3 describes a side-project on copy number variation detection in large biobank data base.

Part 1: Although scRNA-seq is now ubiquitously adopted in studies of intratumor heterogeneity, detection of somatic mutations and inference of clonal membership from scRNA-seq is currently unreliable. We propose DENDRO, an analysis method for scRNA-seq data that detects genetically distinct subclones, assigns each single cell to a subclone, and reconstructs the phylogenetic tree describing the tumor's evolutionary history. DENDRO utilizes information from single nucleotide mutations in transcribed regions and accounts for technical noise and expression stochasticity at the single cell level. The accuracy of DENDRO was benchmarked on spike-in datasets and on scRNA-seq data with known subpopulation structure. We applied DENDRO to delineate subclonal expansion in a mouse melanoma model in response to immunotherapy, highlighting the role of neoantigens in treatment response. We also applied DENDRO to primary and lymph-node metastasis samples in breast cancer, where the new approach allowed us to better understand the relationship between genetic and transcriptomic intratumor variation.

Part 2: Recent technological advances allow the simultaneous profiling, across many cells in parallel, of multiple omics features in the same cell. In particular, high throughput quantification of the transcriptome and a selected panel of cell surface proteins in the same cell is now feasible through the REAP-seq and CITE-seq protocols. Yet, due to technological barriers and cost considerations, most single cell studies, including Human Cell Atlas (HCA) project, quantify the

transcriptome only and do not have cell-matched measurements of relevant surface proteins that can serve as integral markers of cellular function and targets for therapeutic intervention. Here we propose cTP-net (single cell Transcriptome to Protein prediction with deep neural network), a transfer learning approach based on deep neural networks, that imputes surface protein abundances for scRNA-seq data. Through comprehensive benchmark evaluations and applications to HCA and AML data sets, we show that cTP-net outperform existing methods and can transfer information from training data to accurately impute 24 immunophenotype markers, which achieve a more detailed characterization of cellular state and cellular phenotypes than transcriptome measurements alone. cTP-net relies, for model training, on accumulating public data of cells with paired transcriptome and surface protein measurements.

Part 3: Copy number variations (CNVs) are gains and losses of DNA segments that are highly associated with multiple diseases. The Penn Medicine BioBank stores SNP-array and NGS data for more than 10000 individuals across ethnicity and conditions, providing a rich resource for CNV discovery and analysis. This type of experiment design fits perfectly for CNV detection tool - Integrated Copy Number Variation caller (iCNV), which I developed as my master thesis. The distinguishing feature of iCNV includes adaptation of platform specific normalization, utilization of allele specific reads from sequencing and integration of matched NGS and SNP-array data by a Hidden Markov Model (HMM). We applied iCNV on Penn Medicine BioBank data set, calling CNV over more than 10000 individuals (~2000 AFR, ~8000 EUR) with different phenotypes. iCNV detected on average 34.1 deletions and 11.3 duplications per EUR sample, and 38 deletions and 10.6 duplications per AFR sample. iCNV calling results show great improvement in detection sensitivity and specificity comparing to single platform detection method. Penn Medicine BioBank CNV sets by iCNV provide a rich database for researchers to study the relationship between diseases phenotypes and CNV across ethnicity and conditions.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1 DENDRO: GENETIC HETEROGENEITY PROFILING AND SUBCLONE DETECTION BY SINGLE-CELL RNA SEQUENCING

1.1 Introduction

DNA alterations, especially single nucleotide alteration (SNA) and epigenetic modulation both contribute to intratumor heterogeneity [1], which mediates tumor initiation, progression, metastasis and relapse [2, 3]. Intratumor genetic and transcriptomic variation underlie patients' response to treatment, as natural selection can lead to the emergence of subclones that are drug resistant [4]. Thus, identifying subclonal DNA alterations and assessing their impact on intratumor transcriptional dynamics can elucidate the mechanisms of tumor evolution and, further, uncover potential targets for therapy. To characterize intratumor genetic heterogeneity, most prior studies have used bulk tumor DNA sequencing [5-12], but these approaches have limited resolution and power [13].

Breakthroughs in single-cell genomics promise to reshape cancer research by allowing comprehensive cell type classification and rare subclone identification. For example, in breast cancer, single-cell DNA sequencing (scDNA-seq) was used to distinguish normal cells from malignant cells, the latter of which were further classified into subclones [14-16]. For the profiling of intra-tumor transcriptional heterogeneity, single cell RNA-sequencing (scRNA-seq), such as Smart-seq2 [17], Drop-seq [18], and 10X Genomics Chromium™, is now ubiquitously adopted in ongoing and planned cancer studies. ScRNA-seq studies have already led to novel insights into cancer progression and metastasis, as well as into tumor prognosis and treatment response, especially response variability in immune checkpoint blockade (ICB) [19-26]. Characterization of intratumor genetic heterogeneity and identification of subclones using scRNA-seq is challenging, as SNAs derived from scRNA-seq reads are extremely noisy and most studies have relied on the detection of chromosome-level copy number aberrations through smoothed gene expression

profiles. Yet, as intratumor transcriptomic variation is partially driven by intratumor genetic variation, the classification of cells into subclones and the characterization of each subclone's genetic alterations should ideally be an integral step in any scRNA-seq analysis.

The appeal of subclone identification in scRNA-seq data is compounded by the shortage of technology for sequencing the DNA and RNA molecules in the *same* cell with acceptable accuracy, throughput, and cost [27-30]. Although one can apply both scDNA-seq and scRNA-seq to a given cell population, the mutation analysis and RNA quantification cannot be conducted in the same set of cells. Although there are now technologies for deep targeted sequencing of select transcripts matched with same-cell whole transcriptome sequencing [31, 32], these methods are still, in effect, profiling DNA-level variation by sequencing expressed transcripts, and are thus subject to the technical issues, especially dropout due to transcriptional stochasticity.

Subclone detection using scRNA-seq is difficult mainly because only a small portion of the SNAs of each cell is expected to be seen in the read output of scRNA-seq. This is because to be sequenced, an SNA needs to fall in a transcribed region of the genome, at a location within the transcript that will eventually be read by the chosen sequencing protocol. Even for SNAs that satisfy these requirements, the mutated allele are often missing in the read output due to *dropout*, especially in the heterozygous case. This is due, in part, to the bursty nature of gene transcription in single cells [33-35], where in any given cell, a substantial fraction of the genes are only expressed from one of the alleles. Thus, an SNA residing in a gene that is expressed at the bulk tissue level may not be observed in a particular cell, simply because the mutated allele, by chance, is not expressed in the given cell. We refer to alleles that are not captured due to expression stochasticity as *biological dropouts*. Even for a mutated allele that is expressed, it has to be successfully converted to cDNA and then sequenced to be represented in the final read output; we refer to alleles lost due to technical reasons as *technical dropouts*. In addition to dropout events, post-transcriptional modification, such as RNA editing, and sequencing errors

impede both the sensitivity and the specificity of SNA discovery. As a result, methods developed for single cell SNA detection using scDNA-seq, such as Monovar [36], as well as methods designed for SNA detection in bulk DNA or RNA sequencing data do not yield accurate results in the scRNA-seq setting [37-42].

Here we present a new statistical and computational framework – DNA based EvolutionNary tree preDiction by scRNA-seq technOlogy (DENDRO) - that reconstructs the phylogenetic tree for cells sequenced by scRNA-seq based on genetic divergence calculated from DNA-level mutations. DENDRO assigns each cell to a leaf in the tree representing a subclone, and, for each subclone, infers its mutation profile. DENDRO can detect genetically divergent subclones by addressing challenges unique to scRNA-seq, including transcriptional variation and technical noise. A DENDRO clustering of scRNA-seq data allows joint genetic and transcriptomic analysis on the same set of cells.

We evaluate DENDRO against existing approaches, through simulation data sets and a metastasized renal cell carcinoma dataset with known subpopulation labels, and show that DENDRO improved the accuracy of subclone detection. We then demonstrate the DENDRO to biological discovery through two applications. The first application profiles the treatment response in a melanoma model to immune checkpoint blockade therapy. DENDRO identified a subclone that contracted consistently in response to ICB therapy, and revealed that the contraction was driven by the high mutation burden and increased availability of predicted neoantigens. Transcriptional divergence between the subclones in this model was very weak, and thus the neoantigen-driven sub-clonal dynamics would not have been detected without extracting DNA-level information. In the second application to a breast tumor dataset, DENDRO detected subclones and allowed for the joint characterization of transcriptomic and genetic divergence between cells in lymph-node metastasis and cells in primary resections.

The DENDRO package, implemented in R, is available at https://github.com/zhouzilu/DENDRO, where we also provide a power calculation toolkit, DENDROplan, to aid in the design of scRNA-seq experiments for subclonal mutation analysis using DENDRO.

1.2 Results

1.2.1 Method overview

1.2.1.1 Overview of DENDRO model and pipeline

Fig. 1.1a shows an overview of DENDRO's analysis pipeline. Per cell counts of total read coverage (N matrix) and mutation allele read coverage (X matrix) at SNA locations are extracted after read alignment and SNA detection (details in Methods, Fig. 1.2). Based on these matrices, DENDRO then computes a cell-to-cell genetic divergence matrix, where entry (c, c') of the matrix is a measure of the genetic divergence between cells c and c'. Details of this genetic divergence evaluation will be given in the next section. DENDRO then clusters the cells into genetically distinct subclones based on this pairwise divergence matrix, and selects the number of subclones based on inspection of the intra-cluster divergence curve. Reads from the same subclone are then pooled together, and the SNA profile for each subclone is re-estimated based on the pooled reads, which improves upon the previous SNA profiles computed at the single cell level. Finally, DENDRO generates a parsimony tree using the subclone-level mutation profiles to more accurately reflect the evolutionary relationship between the subclones.

1.2.1.2 Genetic divergence evaluation

Due to the high rates of biological and technical dropout, SNA detection within each individual cell lacks sensitivity. We also expect low specificity due to the high base error rate in scRNA-seq protocols. Thus, simple distance measures such as the Hamming or Euclidean distances

evaluated on the raw SNA genotype matrix or the raw allele frequency matrix do not accurately reflect the genetic divergence between cells.

To more accurately estimate the cell-to-cell genetic divergence, we have developed a statistical model that accounts for technical dropout, sequencing error and expression stochasticity. Consider two cells, c and c', and let I_c and $I_{c'}$ index the clonal group to which the cells belong. That is, $I_c = I_{c'}$ if cells c and c' come from the same subclone and thus share the same SNA profile. Let $X_c = (X_{c1}, \dots, X_{cm})$ be the mutation allele read counts for this cell at the m SNA sites profiled, and $N_c = (N_{c1}, \dots, N_{cm})$ be the total read counts at these sites. We define the genetic divergence between the two cells as

$$d_{cc'} = -\log \frac{P(X_c, X_{c'} | N_c, N_{c'}, I_c = I_{c'})}{P(X_c, X_{c'} | N_c, N_{c'})} = \sum_{g=1}^{m} d_{cc'}^{g}$$

$$\text{where } d_{cc'}^{g} = -\log \frac{P(X_{cg}, X_{c'g} | N_{cg}, N_{c'g}, I_c = I_{c'})}{P(X_{cg}, X_{c'g} | N_{cg}, N_{c'g})}.$$

In other words, $d_{cc'}$ is the negative log likelihood of the mutation allele counts of cells c and c', given the total read counts and the event that the two cells belong to the same subclone. If c and c' have mutations in mismatched positions, this likelihood for X_c, $X_{c'}$ conditioned on $I_c = I_{c'}$ would be small, giving a large value for $d_{cc'}$. By the assumption of independence between sites, $d_{cc'}$ is the sum of $d_{cc'}^{g}$, where $d_{cc'}^{g}$ is the contribution of mutation site g to the divergence measure. In characterizing the conditional distribution for X_{cg} and $X_{c'g}$, we use a Beta-Binomial distribution to model expression stochasticity and a Binomial model to capture sequencing errors and rare RNA-editing events. Referring to Fig. 1.1b, mutations residing in bursty genes, such as gene g, would tend to have U-shaped allele frequency distributions and are more likely to be "dropped" due to low or zero expression. In contrary, mutations residing in constitutive (non-bursty) genes, such as gene g' in Fig. 1.1b, would have bell-shaped allele frequency distributions and can be

genotyped more reliably. Thus, even if the read counts for the mutation loci residing in genes g and g' are identical across two cells (c_1 and c_2 in Fig. 1.1c), the locus in g' would contribute a higher value, compared to the locus in g, to the divergence between cells c_1 and c_2. Please see Methods for details.

1.2.2 Accuracy assessment

1.2.2.1 Accuracy assessment by simulation experiment

First, we designed a simulation procedure to assess the accuracy of DENDRO versus existing approaches and to make realistic power projections for subclone detection (Fig. 1.3a). Since DENDRO is currently the only method for SNA-based subclone detection using scRNA-seq data alone, we benchmarked against more straightforward approaches such as hierarchical clustering based on mutation allele frequencies and genotypes respectively. The simulation procedure starts with an assumed evolutionary tree, where the leaves are subclones and mutations can be placed on the branches. In the absence of prior information, a simple tree structure is used, such as the one shown in Fig. 1.3a. Parameters of simulation are (1) total number of mutations, (2) total number of cells, (3) the proportion of cells in each clade, (4) the proportion of mutations along each branch, and (5) mean read coverage across loci. Some of these parameters can be determined using bulk DNA-seq and/or bulk RNA-seq data if available (Methods). Parameters (1-4) determine the mutation profile matrix (Fig. 1.3a). To get the matrix of alternative allele (X_{cg}) and total read counts (N_{cg}) for each mutation loci in each cell, we overlay a reference scRNA-seq data with allele-specific read counts onto a designed mutation matrix, which is generated from the simulated tree (See Methods for details). This allows the simulated datasets to retain the expression stochasticity and sequencing error of real scRNA-seq data. DENDRO is then applied to the read count matrices to obtain the subclone clusters, which is then compared with the known labels. Accuracy is evaluated by three metrics: adjusted Rand index, capture rate and purity (See DENDROplan evaluation metrics in Methods). Such simulation procedure can also

facilitate experiment design, as it predicts the expected clustering accuracy by DENDRO given sequencing parameters and available bulk data for the tumor (See DENDROplan in Methods).

Using the above framework, we conducted a systematic evaluation of DENDRO's subclone detection accuracy on an example scRNA-seq dataset with allelic information [43]. The results, compiled in Fig. 1.3b shows that DENDRO has better performance than simply clustering on mutation allele frequencies or the directly estimated mutation profiles from scRNA-seq data. Due to high burstness of the scRNA-seq dataset and limited sequencing depth, we found that Z-matrix, on average, underperformed in all scenario, indicating the necessity of the DENDRO framework. We also quantified how accuracy depends on the mutation burden, mutation read depth, mutation distribution, subclone cell proportion, and cell populations (Fig. 1.4 and See Methods). Even when there are only 100 mutations with relatively low average coverage (read depth equals to 1), DENDRO can still extract meaningful clustering results (average ARI $\approx$ 0.8). More importantly, variation in total expression of genes does not influence DENDRO's divergence measure. DENDRO shows consistent results in simulation analysis between populations of single cell type and multiple cell types (Fig. 1.4). This is due to DENDRO's reliance only on the distribution of the mutation allele frequency conditioned on the total read coverage, as illustrated by the simulation study (Fig. 1.5). The divergence evaluation reflects solely genetic distance not transcriptomic difference, allowing for easy interpretation.

1.2.2.2 Accuracy assessment on a renal cell carcinoma and its metastasis

We also benchmarked DENDRO against existing methods on the renal cell carcinoma dataset from Kim et al [21] (Fig. 1.3). This dataset contained 116 cells sequenced using the Smart-seq technology [17], obtained from three tumors derived from one patient: a patient-derived xenograft (PDX) from the primary renal cell carcinoma (PDX_pRCC), a biopsy of the metastasis to the lung 1 year after treatment of primary site (Pt_mRCC), and a PDX of the lung metastasis renal cell carcinoma (PDX_mRCC) (Fig. 1.6a). The cells should share common early driver mutations due

to their shared origin from the same patient, but the metastasis and the cultivation of each tumor in separate medium (human or mouse) should have allowed for the accumulation of new mutations. Thus, we expect the three tumors to be clonally distinct. This knowledge allows us to use this dataset to benchmark accuracy and to illustrate how DENDRO enables joint analysis of the genetic and transcriptomic heterogeneity at single cell resolution.

GATK detected 2,867,029 mutation sites across all cells [1]. Mutations that are detected in less than 5% (too rare) or more than 95% (too common) of the cells were removed, which leaves 72,206 mutations. On average, 10801 mutations are detected in each cell and 17.35 cells possess the same mutation for each loci (Fig. 1.6b, c). For majority sites, only few cells have nonzero read coverage, highlighting the fact that many mutations are missed due to technical and biological dropout (Fig. 1.6d) [2-6].

We compared 4 different clustering methods: (1) DENDRO, (2) hierarchical clustering based on the primary genotype matrix Z generated by GATK ($Z_{cg} = 1$ when a mutation g is detected for cell c, $Z_{cg} = 0$ otherwise), (3) hierarchical clustering based on the $\frac{X}{N}$ matrix that preserve the variant allele frequency information and (4) hierarchical clustering based on gene expression ($\log TPM$). DENDRO gives the cleanest separation between the three populations with adjusted Rand Index of 0.932 (1.0 indicates perfect clustering, Fig. 1.3c panel 1), as compared to 0.754 for Z matrix (Fig. 1.3c panel 2), 0.519 for $\frac{X}{N}$ matrix (Fig. 1.3c panel 3) and 0.489 for expression (Fig. 1.3c panel 4). Inspection of the tree shows that, as expected, divergence between primary tumor and metastasis exceeds divergence between patient sample and PDX sample, as PDX_mRCC clusters with Pt_mRCC rather than PDX_pRCC. All of the other three methods successfully separated the primary sample from the metastatic samples, but could not differentiate between the two metastasis samples.

For DENDRO, the intra-cluster divergence curve flattened at 3, and thus we stopped splitting at 3 clusters (Fig. 1.6e and Methods). We annotated the clusters as PDX_mRCC, PDX_pRCC and Pt_mRCC by their cell compositions (Table 1.1a). DENDRO found minimal sharing of subclones among the tumors derived from three sources, and low genetic heterogeneity within each tumor. This is unsurprising since relapsed metastasis consists of cells that have already undergone selection, and since the PDX tumors are each seeded by a small subsample of cells from the original tumor, each tumor consists of unique subclones not detected in other sites [44-46].

DENDRO enables simultaneous clonal assignment and transcriptomic profiling of the same set of cells. Plot of smoothed expression ordered by DENDRO shows unique expression patterns within each subclone (Fig. 1.7). We focused on the comparison of the two metastasized cell populations (metastasis to lung and patient derived mouse xenograph). Even though PDX_mRCC was derived from Pt_mRCC, the DENDRO analysis found substantial genetic divergence between the two cell populations. To investigate further, we performed a differential expression analysis between PDX_mRCC and Pt_mRCC with scDD and MAST, detecting 74 significant differentially expressed genes (Methods, Fig. 1.8e, Table 1.2) [44-46]. Gene ontology analysis classified these 74 genes into two subgroups: immune-related genes and cancer-related genes (Table 1.3) [47]. Immune-related differentially expressed genes are enriched for the terms TNF-α signaling, complement system and allograft rejection. On the other hand, cancer related differentially expressed genes overlap with the pathways including hypoxia, KRAS signaling, mTORC1 signaling and epithelial mesenchymal transition.

Simultaneously, we compare the mutation profiles of these two subclones. 9521 locus have different mutated allele counts between these two populations and were further annotated by ANNOVAR [48]. After filtering, the preserved variants associated with 24 out of 74 differential expressed genes (Table 1.2). Next, we performed a similar GSEA on variants associated genes

to identify mutation-related pathway [47]. Interestingly, variant annotated genes are enriched in cancer-related pathways, including mitotic spindle, mTORC1 signaling, EMT and hypoxia, overlapping substantially with the cancer-related pathways identified by differential expression analysis; in comparison, none of the differentially expressed genes from immune pathways showed up in this mutated gene analysis (Table 1.4). In another word, cancer-related transcriptomic divergence between PDX_mRCC and Pt_mRCC is driven directly by genetic alterations in the same genes, but immune-related differential expression is influenced by non-DNA factors. This makes sense, since implantation of tumor cells from human to mice alters their immune microenvironment [49-51], and thus is expected to alter immune-related signaling within the implanted tumor cells. This illustrates how DENDRO extricates DNA variation from RNAs allowing their joint analysis. Differential expression and differential mutation analysis for the other subclone pairs can be found in Fig. 1.8.

1.2.3 DENDRO analysis of melanoma model in response to immune checkpoint blockade highlights the role of neoantigens

Immune checkpoint blockade (ICB) of the inhibitory receptors CTLA4 and PD1 can result in durable responses in multiple cancer types [47]. Features intrinsic to cancer cells that can impact ICB treatment outcome include their repertoire of neoantigens [48], tumor mutational burden (TMB) [49], and expression of PDL1 [50]. DENDRO analysis of scRNA-seq data allows joint DNA-RNA analysis of single cells, thus enabling the simultaneous quantification of tumor mutational burden, the prediction of neoantigen repertoire, and the characterization of gene expression profile at subclonal resolution. Thus, to demonstrate the power of DENDRO and to better understand the relationship between ICB response and intratumor heterogeneity, we profiled the single cell transcriptomes across three conditions derived from 2 melanoma cell lines (Fig. 1.9a): B16 melanoma cell line, which has shown modest initial response to ICB treatment but eventually grows out, and Res 499 melanoma cell line (R499), which was derived from a

relapsed B16 tumor after combined treatment of radiation and anti-CTLA4 and is fully resistant to ICB [51]. B16 was evaluated with and without anti-PD1 treatment, as we wanted a tumor model that captures a transient ICB response. A total of 600 tumor cells were sequenced with Smart-seq technology from six mice across three conditions: two mice with B16 without treatment (B16), two mice with B16 after anti-PD1 treatment (B16PD1) and two mice with R499 without treatment (R499) (Fig. 1.9a and Methods). The existence of multiple subclones in B16 and R499 was suggested by bulk WES analysis [51, 52]. Our goal here is to determine whether the subclones differ in anti-PD1 response, and if so, what are the subclonal differences.

A DENDRO analysis of 4059 putative mutation sites across 460 cells retained after QC (see Methods and Fig. 1.10a, b, c) yields the clustering displayed in Fig. 1.9b, with four subclones suggested by the intra-cluster divergence curve (Fig. 1.10d). All subclones are shared among the three conditions, which is not unexpected given that all tumor cells were derived from the same parental cell line. However, the sub-clonal proportions vary significantly between conditions (Fig. 1.9b). The subclonal proportions of B16PD1 are approximately intermediate between that of B16 and R499 (Fig. 1.9c). This is expected as R499 had gone through immune editing whereas B16PD1, at the time of harvest, was still undergoing immune editing and was at the transient response state. Furthermore, the selective pressure of radiation plus anti-CTLA4 is likely more than that of anti-PD1 treatment, as the former but not the latter results in complete responses in our B16 model [51]. The frequency of Clone 2 is lower in B16PD1 and R499, indicating sensitivity to anti-PD1 treatment, while the frequencies of Clone 3 and Clone 4 increase after treatment and are the highest in R499, indicating resistance to therapy (Fig. 1.9c, 1.11a).

To explore why subclones vary in sensitivity to anti-PD1 treatment, we compared the mutation profile of Clone 2 to the other subclones. We pooled cells in each of the four subclones and re-estimated their mutation profiles, which were then used to construct a phylogenetic tree (Fig. 1.9d). The phylogeny suggests that Clone 3 and Clone 4 are genetically closer to each other

than to Clone 2, and thus, their similarity in treatment response may be in part due to similarity in their mutation profiles. The re-estimated mutation profiles show that Clone 2 has the highest tumor mutation burden, which has been associated with increased likelihood of ICB response [53, 54]. We then predicted the quantity of high-affinity ($\leq$ 100 nm) neoantigens in each subclone given its mutation profile [52]. As shown in Fig. 1.9e, Clone 2 has twice as many high-affinity neoantigens as the other three subclones. The high level of neoantigens can lead to better T cell recognition, resulting in increased efficacy of anti-PD1 treatment [55].

Analysis of gene expression, on the other hand, did not yield detectable known signatures associated with anti-PD1 treatment sensitivity. Projections based on the expression of highly variable genes, as shown in PCA and t-SNE plots (Fig. 1.12), did not yield meaningful clusters. Differential expression analysis between each subclone and the other subclones found few genes with adjusted P-value < 0.05, indicating similar expression across sub-clones that is concordant with the lack of structure in the expression PCA and tSNE plots. Expressions of *Pdl1* (aka. *Cd274*) showed no differences between subclones (KS-test: P-value > 0.42, Fig. 1.11b). In addition, there were no detectable chromosome-level differences in smoothed gene expression, indicating that there are no large CNV events that distinguish the subclones (Fig. 1.13). DENDRO, detecting exonic mutations from scRNA-seq data, enabled the finding of subclones in this data, the prediction of neoantigen load of each subclone, and the analysis of subclonal dynamics due to treatment. Our analysis suggests that the genetic heterogeneity, rather than transcriptomic heterogeneity, contributes to treatment efficacy in this tumor model.

1.2.4 Simultaneous analysis of genetic and transcriptomic variation in single cell breast cancer

We next applied DENDRO to the analysis of data from a study of primary and metastasized breast cancer [20]. We focused on tumors from two patients (BC03 and BC09) that had the most cells sequenced (Fig. 1.14 and Table 1.5). Patient BC03 had cells sequenced from the primary

tumor (here after BC03P) as well as cells from regional metastatic lymph nodes (here after BC03LN), whereas patient BC09 had cells sequenced only from the primary resection. 132 single cell transcriptomes were profiled by Smart-seq protocol [17]. We first assess whether DENDRO separated BC03 cells from BC09 cells, since inter-individual genetic distances should far exceed intra-individual genetic distances owning to the randomness of passenger mutations [19, 22, 56] . Then, we examine the transcriptomic and genetic heterogeneity within each tumor.

GATK [57] detected a total of 2,364,823 mutation sites across the 132 cells, 353,647 passed QC (Methods) and were retained for downstream analysis (Fig. 1.14a, b, c). Fig. 1.15 shows the clustering determined by DENDRO. DENDRO separates BC09 cells from BC03 cells with 100% accuracy (Fig. 1.15a). The intra-cluster divergence curve flattened at five subclones: three subclones for BC03 and two for BC09 (Fig. 1.15a, Fig. 1.14d and Table 1.1b). Within BC03, Clone Mix_1 and Clone Mix_2 contained a mixture of cells from the primary tumor and lymph nodes, and Clone LN_1 contained mostly cells from the lymph nodes. This suggests that tumor cells that have metastasized to the lymph nodes belong to an intermediate stage and are genetically heterogeneous, with some cells remaining genetically similar to the primary population and others acquiring new genetic mutations, coherent with previous studies [58, 59]. In comparison, hierarchical clustering based on expression (using log transcripts-per-million values) did not separate BC03 from BC09, and gave a negative adjusted Rand index within BC03, indicating effectively random assignment of cells to the two patients (Fig. 1.15b).

We then pooled cells within each of the 5 clusters and re-estimated their mutation profiles with DENDRO. We defined a variant as subclonal if it was not present in all of the subclones within a tumor. Based on detection marginal likelihood, we picked the top 10,000 most confident variants to construct a phylogenetic tree (Fig. 1.15c). As expected, the two BC09 clusters are far from the three BC03 clusters. Within BC03, the length of the branches shows that the subclone containing mostly cells from lymph nodes (labeled BC03LN_1) is genetically more similar to

Clone Mix_2 compared to Clone Mix_1 (Fig. 1.15c). In addition, window-smoothed expression plot with cells grouped by DENDRO clustering shows broad chromosome-level shifts in expression patterns between subclones, most likely due to copy number aberrations that are consistent with SNAs (Fig. 1.16) [22].

A comparison of the transcriptomes of the subclones revealed substantial differences in the expression of PAM50 genes, which are prognostic markers for breast cancer (Fig. 1.15d) [60]. DENDRO detected one rare subclone, BC09_2, with only six cells (<5% of the total number of cells) which had a strong basal-like signature. Interestingly, in BC03, Clone LN_1 has the TNBC/basal-like subtype with an invasive gene signature, while Clone Mix_2 has the *ESR1*+ subtype. Thus, the genetic divergence of Clone LN_1 from Clone Mix_2 is accompanied by its acquisition of an invasive metastatic expression signature. In a direct comparison between cells from the primary site and cells from the lymph node without distinguishing subclones, these expression differences would be much weaker since the subclones do not cleanly separate by site. Compared with the original analysis that assigned each tumor to one specific breast cancer subtype, this analysis identifies subclones with different expression phenotypes, potentially allowing for better therapy design that targets all subclone phenotypes to reduce the risk of tumor relapse.

Existing scRNA-seq studies of cancer tissue cluster cells based on total gene expression or copy number profiles derived from smoothed total expression, making it difficult to separate the effects of sub-clonal copy number aberrations from transcriptomic variation [19, 22, 24]. Differential expression analysis based on clusters derived from total expression is prone to self-fulfilling prophecy, as there would indeed be differentially expressed genes because this is the clustering criteria. Because DENDRO's subclone identification is based solely on genetic divergence, and not on expression profile, the downstream differential gene expression analysis can be precisely attributed to transcriptional divergence between subclones.

Hence, we conducted a transcriptome-wide search for pathways that have differential expression between subclones (Methods and Table 1.6), and assessed their overlap with pathways that are differentially mutated between subclones. Focusing on tumor BC03, pathways for G2M checkpoint and *KRAS* signaling are up-regulated in lymph node metastasis Clone BC03LN_1, while pathways for estrogen response and apoptosis are down-regulated, indicating a more invasive phenotype. In addition, *GAPDH* is up-regulated in the metastatic subclone (BC03LN_1) and down-regulated in the two mix-cell subclones, consistent with previous findings [61, 62] (Fig. 1.17d). Differentially expressed genes between other subclone pairs in BC03 are also enriched in estrogen response, apoptosis, and DNA repair. In parallel, subclone-specific mutated genes are highly enriched in cancer-related pathways including MYC target, G2M checkpoints and mitotic spindle, and immune related pathways such as, interferon response, TNF-a signaling and inflammatory response (Table 1.6). Interestingly, few of the differentially mutated genes are associated with estrogen and androgen responses, suggesting that the differential expression of hormone related genes is not mediated directly by genetic mutations in these pathways. This is consistent with the recent studies that epigenetic alteration, such as histone acetylation and methylation, regulate hormones receptor signaling in breast cancer [63-66]. DNA-RNA joint analysis between other subclones are included in Fig. 1.17. Overall, this example illustrates how DENDRO enables the joint assessment of genetic and transcriptomic contributions to clonal diversity at single-cell resolution.

1.3 Discussion

We have described DENDRO, a statistical framework to reconstruct intratumor DNA-level heterogeneity using scRNA-seq data. DENDRO starts with mutations detected directly from the scRNA-seq reads, which are very noisy due to a combination of factors: (1) errors are introduced in reverse-transcription, sequencing and mapping, (2) low sequencing depth and low molecule conversion efficiency leading to technical dropouts, and (3) expression burstiness at the single cell level leading to biological dropouts. DENDRO overcomes these obstacles through the

statistical modeling of each component. Given noisy mutation profiles and allele-specific read counts, DENDRO computes a distance between each pair of cells that quantifies their genetic divergence after accounting for transcriptional bursting, dropout and sequencing error. Then, DENDRO clusters the cells based on this distance as subclone and re-estimates a more robust subclone-specific mutation profile by pooling reads across cells within the same cluster. These re-estimated mutations profiles are then passed to downstream mutation analysis and phylogenetic tree reconstruction.

Importantly, the genetic divergence used by DENDRO for cell clustering is based solely on allelic expression ratios and do not reflect the difference in total expression between cells at mutation sites. Thus, DENDRO differs from, and complements, existing tools that cluster cells based on total expression. In fact, as shown by simulation analysis, DENDRO clusters the cells based on true underlining mutation profiles, and is robust to changes in total gene expression. As expected, the numbers of cells, the depth of sequencing, the actual number of subclonal mutations and the phylogenetic tree structure all influence the power of DENDRO. To aid researchers in experiment design, we developed DENDROplan, which predicts DENDRO's clustering accuracy given basic experimental parameters and the expected informative mutation count, which can be obtained from bulk DNA sequencing.

Ideally, joint sequencing of the DNA and RNA on the same cells would allow us to relate genomic profiles to transcriptomic variations. Currently, there is yet no scalable technology for doing this. Separately performing scDNA-seq and scRNA-seq on different batches of cells within the same tumor would meet the nontrivial challenge of matching the subclones between the two data sets. DENDRO takes advantage of the central dogma and utilizes computational methods to extract genetic divergence information from noisy mutation calls in coding regions. Through two case studies, we illustrate the insights gained from the subclonal mutation and expression joint analysis that DENDRO enables.

We have demonstrated that proper computational modeling can excavate the DNA-level heterogeneity in scRNA-seq data. Yet, there are always limitations in working with RNA. While rare RNA editing events are absorbed by the parameter ϵ, DENDRO cannot distinguish subclone-specific constituent RNA editing events from subclone-specific DNA mutations. In the extreme and unlikely scenario where RNA editing events are common and pervasive, DENDRO's cluster would reflect RNA editing. In such cases, we recommend using matched bulk DNA-seq of the same tumor to filter the loci detected in the first step of DENDRO, keeping only those that are supported by at least one read in the bulk DNA-seq data. In addition, DENDRO's analysis is restricted to transcribed regions, as variants are detected using transcriptomic data, and thus ignores non-coding mutations which can sometimes be informative for tumor evolution [67-70].

Tag-based scRNA-seq (10X, Drop-seq, etc.) is now commonly adopted for cancer sequencing, but we do not recommend applying DENDRO to this sequencing design because of two reasons: (1) limited number of variants can be detected with tag-based methods as they only profile a small fraction of the transcript (3-prime or 5-prime end); and (2) the sequencing depth of tag-based methods are critically low (<0.1X), resulting in unreliable variant calling. However, we do anticipate that emerging technologies, such as long-read full-transcript scRNA-seq technologies [71] and transcriptome-based deep targeted sequencing [31, 32] will overcome these limitations of tag-based scRNA-seq. Given proper experimental design, we expect that these emerging technologies will be ideally suited for the joint analysis of exonic somatic mutations and gene expression.

1.4 Methods

1.4.1 scRNA-seq alignment and SNA calling pipeline

Fig. 1.2 illustrates the SNA calling pipeline. Raw scRNA-seq data is aligned by STAR 2-pass method (default parameters), which accounts for splicing junctions and achieve higher

mapping quality [72]. Transcripts per million (TPM) was quantified using RSEM (default parameters) [73]. In the next step, raw variants calling is made using the Haplotype Caller (GATK tool) on the BAM files after sorting, joining read groups, removing duplicated reads, removing overhangs into intronic regions, realigning and recalibration [74]. Conventionally, there are two methods from GATK tools for mutation detection: haplotype caller and mutect2. Haplotype caller has a RNA-seq setting which handle splice junctions correctly, but assumes VAF around 50%, while mutect2 can detect mutations with low VAF but does not account for splice junction. The reason we select haplotype caller instead of mutect2 is that we extract allele read counts for all cells as long as one of the cells is listed as carrying the mutation. Thus, as long as one cell has VAF reaching 50%, this mutation would be detected. Calls with stand_call_conf greater than 20 and population frequency greater than 5% but less than 95% were preserved for further analysis. Admittedly, such lenient filtering typically introduces false positive sites. However, our priority at this step is to minimize false negative rate, while the genetic divergence matrix in the following step robustly estimates cell population substructure. Both the coverage of the alternative allele and the total read coverage are extracted for each site for further analysis.

1.4.2 Data preprocessing and quality control

To ensure robustness of downstream analysis, we filtered out low quality cells, variants and genes. We retained: Cells with (1) >10000 reads mapped, (2) <10% mitochondria gene expression and (3) >1000 gene detected; genes with > 5 cells detected (TPM>0 as detected); and variants with > 2 cells detected by GATK. Original TPM values as defined by RSEM were added a value of 1 (to avoid zeros) and then log-transformed for downstream transcriptomic analysis.

1.4.3 Genetic Divergence and Beta-Binomial framework

Consider two cells: c and c'. Let I_c and $I_{c'}$ denote the clonal group to which the cells belong, i.e. $I_c = I_{c'}$ if and only if cells c and c' come from the same subclone. We define the genetic divergence at loci g, by $d_{cc'}^{g}$:

$$d_{cc'}^{g} = \log \frac{\mathrm{P}\left(X_{cg}, X_{c'g} \middle| N_{cg}, N_{c'g}\right)}{P(X_{cg}, X_{c'g} | N_{cg}, N_{c'g}, I_c = I_{c'})}$$

$$= \log \frac{\mathrm{P}\left(X_{cg}, X_{c'g} \middle| N_{cg}, N_{c'g}, I_c = I_{c'}\right) + \mathrm{P}(X_{cg}, X_{c'g} | N_{cg}, N_{c'g}, I_c \neq I_{c'})}{P(X_{cg}, X_{c'g} | N_{cg}, N_{c'g}, I_c = I_{c'})})$$

where $X_c = (X_{c1}, X_{c2}, \dots X_{cg}, \dots X_{cm})$ are the mutation allele read counts for cell c and $N_c = (N_{c1}, N_{c2}, \dots N_{cg}, \dots N_{cm})$ are the total read counts at these sites. More intuitively, if cells c and c' are not from the same clonal group, the probability of cell cells c and c' from the same cells given data (i.e. denominator) has smaller value. Thus $d_{cc'}^{g}$ is large, indicating bigger divergence between the two cells.

Given $d_{cc'}^{g} = -\log \frac{\mathrm{P}\left(X_{cg}, X_{c'g} \middle| N_{cg}, N_{c'g}, I_c = I_{c'}\right)}{\mathrm{P}\left(X_{cg}, X_{c'g} \middle| N_{cg}, N_{c'g}\right)}$,

$$d_{cc'}^{g} = \log \frac{\mathrm{P}\left(X_{cg}, X_{c'g} \middle| N_{cg}, N_{c'g}\right)}{P\left(X_{cg}, X_{c'g} \middle| N_{cg}, N_{c'g}, I_c = I_{c'}\right)}$$

$$= \log \frac{\mathrm{P}\left(X_{cg}, X_{c'g} \middle| N_{cg}, N_{c'g}, I_c = I_{c'}\right) + \mathrm{P}\left(X_{cg}, X_{c'g} \middle| N_{cg}, N_{c'g}, I_c \neq I_{c'}\right)}{P\left(X_{cg}, X_{c'g} \middle| N_{cg}, N_{c'g}, I_c = I_{c'}\right)})$$

$$= \log \left(\frac{P(X_{cg}, X_{c'g} | N_{cg}, N_{c'g}, I_c \neq I_{c'})}{P(X_{cg}, X_{c'g} | N_{cg}, N_{c'g}, I_c = I_{c'})} + 1\right)$$

where $D_c = \{(N_{c1}, N_{c2}, \dots N_{cg}, \dots N_{cm}), (X_{c1}, X_{c2}, \dots X_{cg}, \dots X_{cm})$ are data for cell c.

$\frac{P(X_{cg}, X_{c'g} | N_{cg}, N_{c'g}, I_c \neq I_{c'})}{P(X_{cg}, X_{c'g} | N_{cg}, N_{c'g}, I_c = I_{c'})}$ could also be called a Bayes Factor. More intuitively, if cell c and c' are not

from the same clonal group, the numerator has larger value compared to denominator. Thus, $d_{cc'}^{g}$ is large, indicating bigger divergence between the two cells.

To further expand the formula, let us focus on the denominator first:

$$P\left(X_{cg}, X_{c'g} \,|N_{cg}, N_{c'g}, I_c = I_{c'}\right)$$

$$= P\left(X_{cg}, X_{c'g} \,|N_{cg}, N_{c'g}, Z_{cg} = Z_{c'g} = 0\right)P\left(Z_{cg} = Z_{c'g} = 0 | I_{cg} = I_{c'g}\right)$$
$$+ P\left(X_{cg}, X_{c'g} \,|N_{cg}, N_{c'g}, Z_{cg} = Z_{c'g} = 1\right)P\left(Z_{cg} = Z_{c'g} = 1 | I_{cg} = I_{c'g}\right)$$

$$= P\left(X_{cg} \,|N_{cg},\, Z_{cg} = 0\right)P\left(X_{c'g} \,|N_{c'g},\, Z_{c'g} = 0\right)\left(1 - P_g\right)$$
$$+ P\left(X_{cg} \,|N_{cg},\, Z_{cg} = 1\right)P\left(X_{c'g} \,|N_{c'g},\, Z_{c'g} = 1\right)P_g$$

where $P_g = P\left(Z_g = 1\right)$ indicates the population mutation frequency in the group of cells, estimated from GATK calling; and $P\left(Z_g = 0\right) = 1 - P\left(Z_g = 1\right) = 1 - P_g$.

Then the numerator:

$$P\left(X_{cg}, X_{c'g} \,|N_{cg}, N_{c'g}, I_c \neq I_{c'}\right)$$

$$= P\left(X_{cg}, X_{c'g} \,|N_{cg}, N_{c'g}, Z_{cg} = 0,\, Z_{c'g} = 0\right)P\left(Z_{cg} = 0, Z_{c'g} = 0 | I_{cg} \neq I_{c'g}\right)$$
$$+ P\left(X_{cg}, X_{c'g} \,|N_{cg}, N_{c'g}, Z_{cg} = 1,\, Z_{c'g} = 1\right)P\left(Z_{cg} = 1, Z_{c'g} = 1 | I_{cg} \neq I_{c'g}\right)$$
$$+ P\left(X_{cg}, X_{c'g} \,|N_{cg}, N_{c'g}, Z_{cg} = 0,\, Z_{c'g} = 1\right)P\left(Z_{cg} = 0, Z_{c'g} = 1 | I_{cg} \neq I_{c'g}\right)$$
$$+ P\left(X_{cg}, X_{c'g} \,|N_{cg}, N_{c'g}, Z_{cg} = 1,\, Z_{c'g} = 0\right)P\left(Z_{cg} = 1, Z_{c'g} = 0 | I_{cg} \neq I_{c'g}\right)$$

$$= P\left(X_{cg} \,|N_{cg},\, Z_{cg} = 1\right)P\left(X_{c'g} \,|N_{c'g},\, Z_{c'g} = 0\right)\left(1 - P_g\right)P_g$$
$$+ P\left(X_{cg} \,|N_{cg},\, Z_{cg} = 0\right)P\left(X_{c'g} \,|N_{c'g},\, Z_{c'g} = 0\right)\left(1 - P_g\right)^2$$
$$+ P\left(X_{cg} \,|N_{cg},\, Z_{cg} = 1\right)P\left(X_{c'g} \,|N_{c'g},\, Z_{c'g} = 1\right)P_g^2$$
$$+ P\left(X_{cg} \,|N_{cg},\, Z_{cg} = 0\right)P\left(X_{c'g} \,|N_{c'g},\, Z_{c'g} = 1\right)\left(1 - P_g\right)P_g$$

As a result, $d_{cc'}^{g}$ is a function of the five following probabilities:

$$d_{cc'}^{g}$$
$$= f\left(P_g; P\left(X_{cg}\middle|N_{cg}, Z_{cg}=0\right); P\left(X_{cg}\middle|N_{cg}, Z_{cg}=1\right); P\left(X_{c'g}\middle|N_{c'g}, Z_{c'g}=0\right); P\left(X_{c'g}\middle|N_{c'g}, Z_{c'g}=1\right)\right)$$

where $Z_{cg} \in \{0,1\}$ is SNA indicator for cell c at site g and $P_g = P(Z_g = 1)$ is mutation frequency across the cells estimated by GATK calls.

In the above formula for $d_{cc'}^{g}$, $P\left(X_{cg}\middle|N_{cg}, Z_{cg}=0\right)$ and $P\left(X_{c'g}\middle|N_{c'g}, Z_{c'g}=0\right)$ reflect reverse-transcription/sequencing/mapping errors and rare RNA editing events, because when there is no mutation (i.e. $Z_{cg}=0, Z_{c'g}=0$), all mutation reads reflect such technical errors or RNA editing. Let ϵ denote the combined rate of technical error and RNA editing, we have

$$P\left(X_{cg}\middle|N_{cg}, Z_{cg}=0\right) \sim \text{Binomial}(X_{cg}|N_{cg}, \epsilon)$$

where ϵ is set to 0.001 based on prior knowledge {Pfeiffer, 2018 #411}.

For cases where there are mutations (i.e. $Z_{cg}=1$), the distribution of mutated read counts given total read counts is modeled with a Beta Binomial distribution, which is capable of modeling technical dropout and transcriptional bursting, and is supported by previous allele specific expression studies [34, 75] .

$$P\left(X_{cg}\middle|N_{cg}, Z_{cg}=1\right) \sim \int_0^1 \text{Binomial}\left(X_{cg}\middle|N_{cg}, Q_{cg}=q\right) dF(q),$$

$$q \sim \text{Beta}(\alpha_g, \beta_g)$$

where Q_{cg} indicates proportion of mutated alleles expressed in cell c at site g, with Beta distribution as prior. Respectively, α_g and β_g represent gene activation and deactivation rate, which are estimated empirically across cells based on first and second moment estimator.

Through optimized vectorization, given a data set of 500 cells with 2500 variants, genetic divergence matrix can be computed under 2 mins in a normal desktop with 16GB of RAM (single thread). Analytically, the algorithm is of complexity $O(N^2 * G)$, where N is number of cells and G is number of variants.

1.4.4 Kernel based clustering and optimal cluster assignment

We cluster the cells using a kernel-based algorithm, such as hierarchical clustering. Given that there are multiple sorting schemes, we leave the user to choose it. For the default-sorting scheme, we recommend "ward.D" [76]. This is because $d_{cc'}$ behaves like a log likelihood ratio, which should follow a χ^2 distribution when the two cells share the same subclone. The "ward.D" method has been shown to work well in Euclidian space. Empirically, among different hierarchical clustering algorithms on the renal cell carcinoma dataset (Fig. 1.18) "ward.D" based hierarchical clustering performs the best.

To determine the number of clusters we use an intra-cluster divergence curve computed from the divergence matrix. Existing software rely on AIC, BIC, or another model selection metric [77, 78]. However, since we only have the "distance" matrix, these traditional methods cannot be applied. Let N_k be the number of cell pairs in cluster C_k and N be the total number of pairs between cells for all clusters. Let K be the number of clusters. The weighted sum of intra-cluster distance W_K is

$$W_K = \sum_{k=1}^{K} N_k \sum_{(i,j)\in C_k} \frac{d_{ij}}{N}$$

Note that small clusters are naturally down-weighted in the above metric. DENDRO relies on visual examination of the intra-cluster divergence curve (W_K plotted against K) to find the "elbow point", which can be taken as a reasonable clustering resolution.

1.4.5 Simulation analysis

In our simulation analysis, we adopt a scRNA-seq dataset from Deng et al. as the reference, which, by crossing two mouse strains, obtained transcriptomic allele specific read counts for every SNPs in exonic regions in each cell [43]. In this case, the Deng et al. data maintained the expression stochasticity in scRNA-seq data. To overlay the read counts on simulated mutation profile, for every simulated locus, we sampled a SNP from this reference. For cells with mutation at this locus, we randomly assigned one allele of the sampled SNP as mutated allele. For cells without mutation, we set the mutated allele counts as 0 and the total read counts as sum of the two alleles from the reference. We further added binomial noise ($p_\epsilon = 0.001$, suggested by [79]) to mimic sequencing error. When analyzing DENDRO performance in terms of various number of mutation sites, number of cells, proportion of cells in each clade and proportion of mutations along each branch, we only take a subset of cells (cells in early blastocyst, mid blastocyst and late blastocyst stages) to ensure the expression homogeneity. On the other hand, we utilize a mixture cell population (cells in 16-cell stages and blastocyst stages) to test the robustness of DENDRO performance with regard to various expression profiles.

1.4.6 Power analysis toolkit and experimental design

Before conducting a single cell RNA-seq experiment on a tumor sample, it is important to project how subclone detection power depends on the number of cells sequenced and the coverage per cell. To facilitate experiment design, we have developed a tool, DENDROplan (Fig. 1.3a), that predicts the expected clustering accuracy by DENDRO given sequencing parameters and available bulk data for the tumor. Given an assumed tree structure and a target accuracy, DENDROplan computes the necessary read depth and number of cells needed.

We evaluate DENDRO accuracy in DENDROplan with three different metrics: Adjusted Rand index, capture rate and purity.

1. Adjusted Rand index: Adjusted Rand index is a measure of the similarity between two data clusterings after adjusted for the chance grouping of elements. For details, see https://en.wikipedia.org/wiki/Rand_index

2. Capture rate: Capture rate is a measure of “false negative rate” of a specific clade. Out of all the cells from the specific clade, how many of them is detected by the algorithm.

3. Purity: Purity is a measure of “false positive rate” of a specific clade. Out of all the cells in the “specific cluster” you detected, how many are actually from the true specific clade.

As shown in Fig. 1.3a, if bulk DNA sequencing and/or RNA sequencing data are available for the tumor being studied, these data can be harnessed to make more realistic power calculations. For example, if SNAs have been profiled using bulk DNA sequencing data, the set of mutations that lie in the exons of annotated genes can be retrieved and used directly in constructing the simulation data. Furthermore, phylogeny construction algorithms for bulk DNA-seq data can be used to infer a putative tree structure that can be used as input to DENDROplan [5, 78]. If bulk RNA-seq data is available, the bulk expression level of the mutation-carrying genes can be used to predict the expression level of the mutation in the single cell data. In another word, variants in high-expressed genes in bulk will be sampled from high-expressed variant loci in scRNA reference and vice versa. The power analysis tool is also available at https://github.com/zhouzilu/DENDRO.

1.4.7 SNA inference in “bulk” and phylogenetic tree construction

As stated previously, DENDRO further inferred SNA after pooling the reads from all cells within each cluster. Because, with our choice of thresholds, we identify SNAs in single cells with high sensitivity, the “bulk” level SNAs should be a subset of the SNAs in single cells, and mutation allele counts and total allele counts should provide us with enough information for SNA detection using a maximum likelihood framework [80], which accounts for both sequencing error and rare

RNA-editing events. Suppose s is the genotype (number of reference allele) at a site and assume m, the ploidy, equals to 2. Then the likelihood is:

$$\mathcal{L}(s) = \frac{1}{m^k} \prod_{j=1}^{l} [(m-s)\epsilon + s(1-\epsilon)] \prod_{j=l+1}^{k} [(m-s)(1-\epsilon) + s\epsilon]$$

where k is number of reads at a site and the first l bases ($l \leq k$) be the same to reference and the rests are same to alternative allele. ϵ is the sequencing error and rare RNA-editing combined rate. s^* is the maximum likelihood estimator of the genotype:

$$s^* = \underset{s}{\operatorname{argmax}} -\mathcal{L}(s)$$

Given mutation profiles, DENDRO then constructs a phylogenetic tree with the neighbor-joining method, which can more accurately capture the evolutionary relationship between different subclones [81] than the initial tree given by hierarchical clustering.

1.4.8 Differential gene expression, mutation annotation and gene ontology analysis

We use Seurat and scDD to identify differentially expressed genes between tumors and between tumor subclones [82-84]. For each comparison, we apply two different methods: MAST implemented by Seurat and scDD. Genes with adjusted p-value < 0.05 count as significant differentially expressed gene for each method. We further intersect these two sets of differentially expressed genes to increase robustness. Subclonal mutations are annotated by ANNOVAR with default parameters and variants associated with intergenic regions were discarded for downstream analysis [85]. For GO analysis, we apply Gene Set Enrichment Analysis tool [57]. Hallmark gene sets serve as fundamental database with FDR q-value < 0.05 as significant.

1.4.9 Single cell RNA-seq of Tumor Model Derived from B16

Six C57bl/6 mice were injected on both flanks with either B16 or R499: four with B16 and two with R499. Two of the mice implanted with B16 were treated with 200 ug of anti-PD1 per mouse on

Days 5, 8 and 11. On Day 15, all tumors were harvested and made into single cell suspension. 100,000 CD45 negative tumor cells were sorted on Aria to enrich for live tumor cells and loaded on SMARTer ICELL8 cx Single-Cell System prior to full length single cell RNA-sequencing library preparation using Smart-seq following manufacturer's recommendations. 460 cells and 11531 genes passed standard QC and were retained for downstream analysis.

1.4.10 Neoantigen prediction

Based on gene expression from RNA-seq data, variants from unexpressed transcripts are removed. The MHC-I binding affinities of variants are then predicted using NetMHC version 4.0 for H-2-Kb and H-2-Db using peptide lengths from 8 to 11 [86]. Given subclonal mutation profile, we further assign the neoantigens to each subclone.

1.4.11 Quantitative function analysis on genetic divergence evaluation by simulation

Better understanding of the genetic divergence evaluation function is essential to DENDRO. Especially, we need to make sure DENDRO is capturing DNA level information from mutation rather than RNA level information from relative expression. This function, however, is quite complicated and difficult to analyze directly. As a result, we design several simulation schemes and analyze the function performance given different variables.

Let's consider 2 cells: cell c and c'. We design true mutation profile Z (an indicator function, 1 means mutation and 0 means no mutation) and relative read count θ (relative number by $\frac{\theta_i}{\sum \theta_j}$) at each position as Fig. 1.5.

We further define true mutation distance d_z as

$$d_z = \frac{|Z_c - Z_{c'}|_1}{m} = p$$

where m is the total number of mutations. In another word, $d_z = p$, which is the true mutation rate in cell c.

We also define $\Delta = |H - M| = |M - L|$, representing relative expression differences. Here, H, M and L represent high expression, medium expression and low expression respectively.

Then, the relative expression distance d_θ can be written as,

$$d_\theta = \sum_i \frac{|\theta_i - \theta_{i'}|_1}{m} = (1-p)\Delta$$

We further have genetic divergence, d_L, calculated by negative log likelihood in DENDRO.

It is also interested in studying the relationship with N, the total number of read counts for each cell.

In our simulation, we alter Δ, p and N, assessing the responses in d_L, d_θ and d_z. Results show that (1) our genetic divergence function is orthogonal to relative expression level (Fig. 1.5b). With fixed mutation rate p and total read counts N, as relative expression differences d_θ increases, d_L stay constant; (2) When we keep relative expression differences d_θ and total read counts N as constant, as number of true mutation (d_z or p) increases, genetic divergence d_L also increases (Fig. 1.5c); and (3) genetic divergence d_L monotonically increases with total expression level (N) if we fix mutation distance d_z (Fig. 1.5d). This can be interpreted that with more reads support, DENDRO is more confident that the two cells belong to different clusters.

1.5 Conclusions

We have developed DENDRO, a statistical method for tumor phylogeny inference and clonal classification using scRNA-seq data. DENDRO accurately infers the phylogeny relating the cells

and assigns each single cell from the scRNA-seq data set to subclone. DENDRO allows us to (1) cluster cells based on genetic divergence while accounting for transcriptional bursting, technical dropout and sequencing error, as benchmarked by in silico mixture and a simulation analysis, (2) characterize the transcribed mutations for each subclone, and (3) perform single-cell multi-omics analysis by examining the relationship between transcriptomic variation and mutation profile with the same set of cells. We evaluate the performance of DENDRO through a simulation analysis and a data set with known subclonal structure. We further illustrate DENDRO through two case studies. In the first case study of relationship between intratumor heterogeneity and ICB treatment response, DENDRO estimates tumor mutation burden and predicts repertoire of high-affinity neoantigens in each subclone from scRNA-seq. In the second case study on a primary breast tumor dataset, DENDRO brought forth new insights on the interplay between intratumor

transcriptomic variation and subclonal divergence.

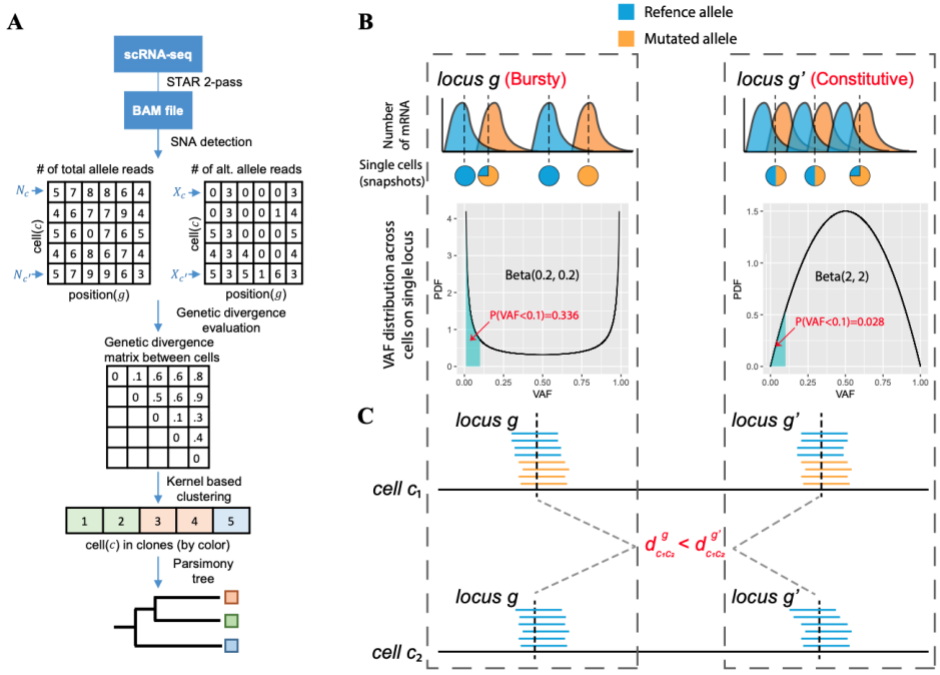


Figure 1.1 ***DENDRO analysis pipeline and genetic divergence evaluation. a*** *DENDRO analysis pipeline overview.* ***b, c*** *Statistical model for genetic divergence evaluation function.* ***b*** *(top) Cell-level snapshots of the variant allele frequency (VAF) profiles for two genes with underlying differences in expression dynamics are shown. Gene* g *is a bursty gene and* g' *is a constitutive gene. (bottom) The stochasticity of gene expression is captured by the VAF distribution across all cells.* ***c*** *Although the observed read counts from two potential cells (*c_1 *and* c_2*) in the population are identical between the two loci, the genetic divergence computed from gene* g *is less than that computed from gene* g' *due to differences in transcriptional burstiness. DENDRO accounts for the full distribution of frequency profiles across cells when estimating the genetic divergence relationship between the two loci of these two cells.*

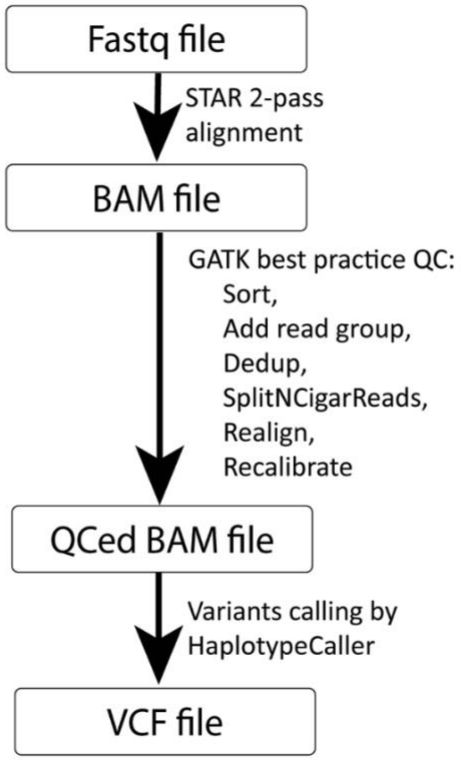


Figure 1.2 ***An illustration of the SNA calling pipeline.*** *Raw scRNA-seq data is aligned by STAR 2-pass. Further quality control and variants calling steps follow GATK tool best practice by Broad Institute*

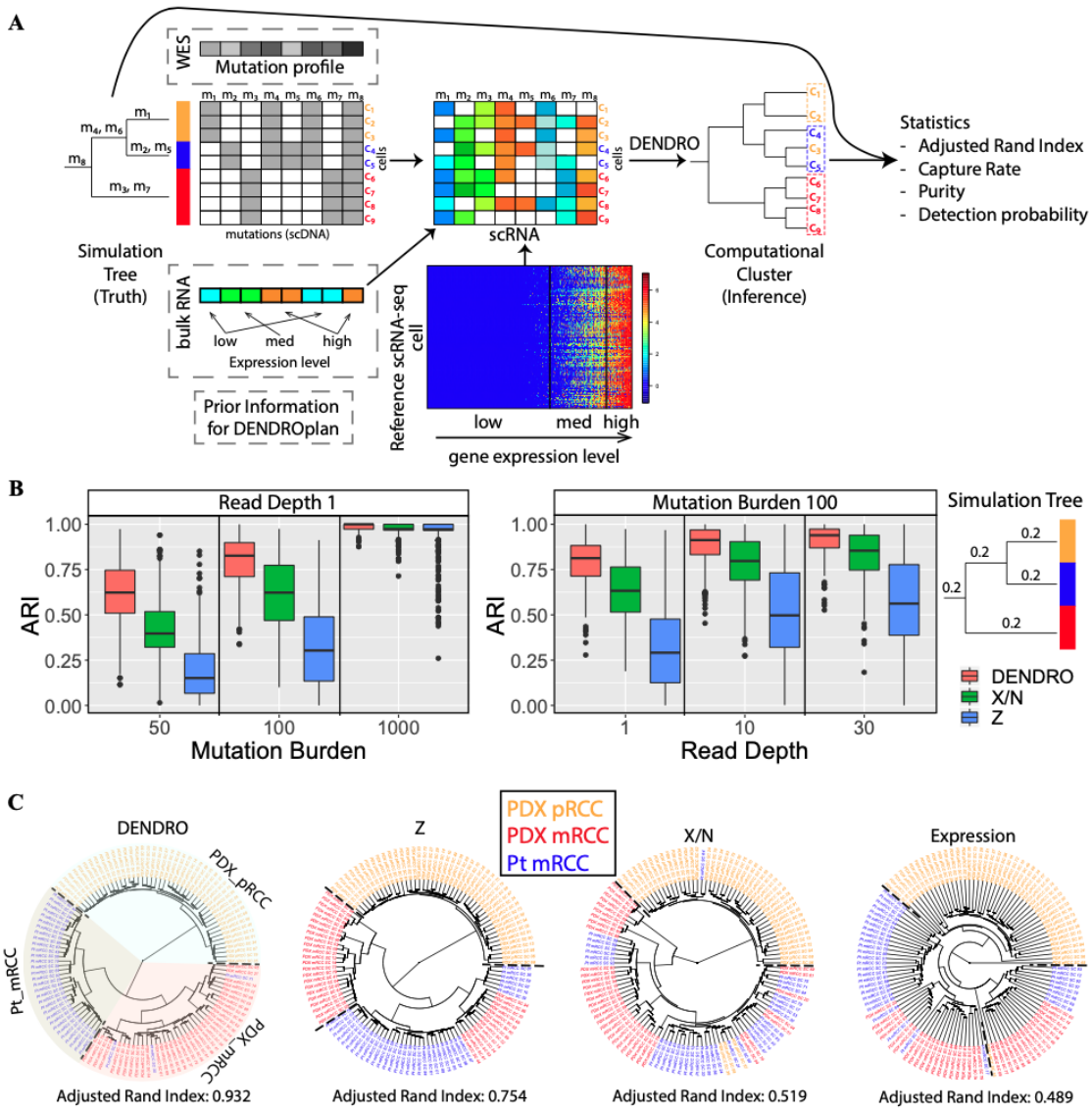


Figure 1.3 ***DENDRO accuracy assessment. a*** *The overall simulation analysis pipeline. Mutation matrix (cell-by-loci) is generated according to a simulated evolutionary tree, where the leaves are subclones and mutations can be placed on the branches. Matrices of alternative allele (X_{cg}) and total read counts (N_{cg}) are sampled from a scRNA-seq dataset with known transcriptomic allele specific read counts. DENDRO cluster is further applied and its performance is assessed by adjusted Rand index (global accuracy), capture rate (subclone-specific sensitivity) and purity (subclone-specific precision). See Methods for detailed definition. Grey dashed line indicates optional input for DENDROplan, where bulk DNA-seq and bulk RNA-seq can guide the tree*

simulation and read count sampling procedure. ***b*** *Cluster accuracy via simulation studies. Various parameters show effects on cluster accuracy (measured by adjusted Rand index) based on tree structure on the most right. Left panel: effect of mutation burden on fixed read depth. Right panel: effect of read depth on fixed mutation burden.* ***c*** *Evaluation of DENDRO on a renal cell carcinoma and its metastasis. (Left to right) (1) DENDRO clustering result from primary and metastatic renal cell carcinoma dataset. Background colors represent DEDRO clustering result. (2) Clustering of the same dataset using* Z *matrix (indicator matrix,* $Z_{ij} = 1$ *when detected a mutation for cell* i *at locus* j *by GATK tool). (3) Clustering of the same dataset using* $\frac{X}{N}$ *matrix (mutation allele frequency matrix) (4) Clustering of the same dataset using expression (*$log\ (TPM + 1)$*).*

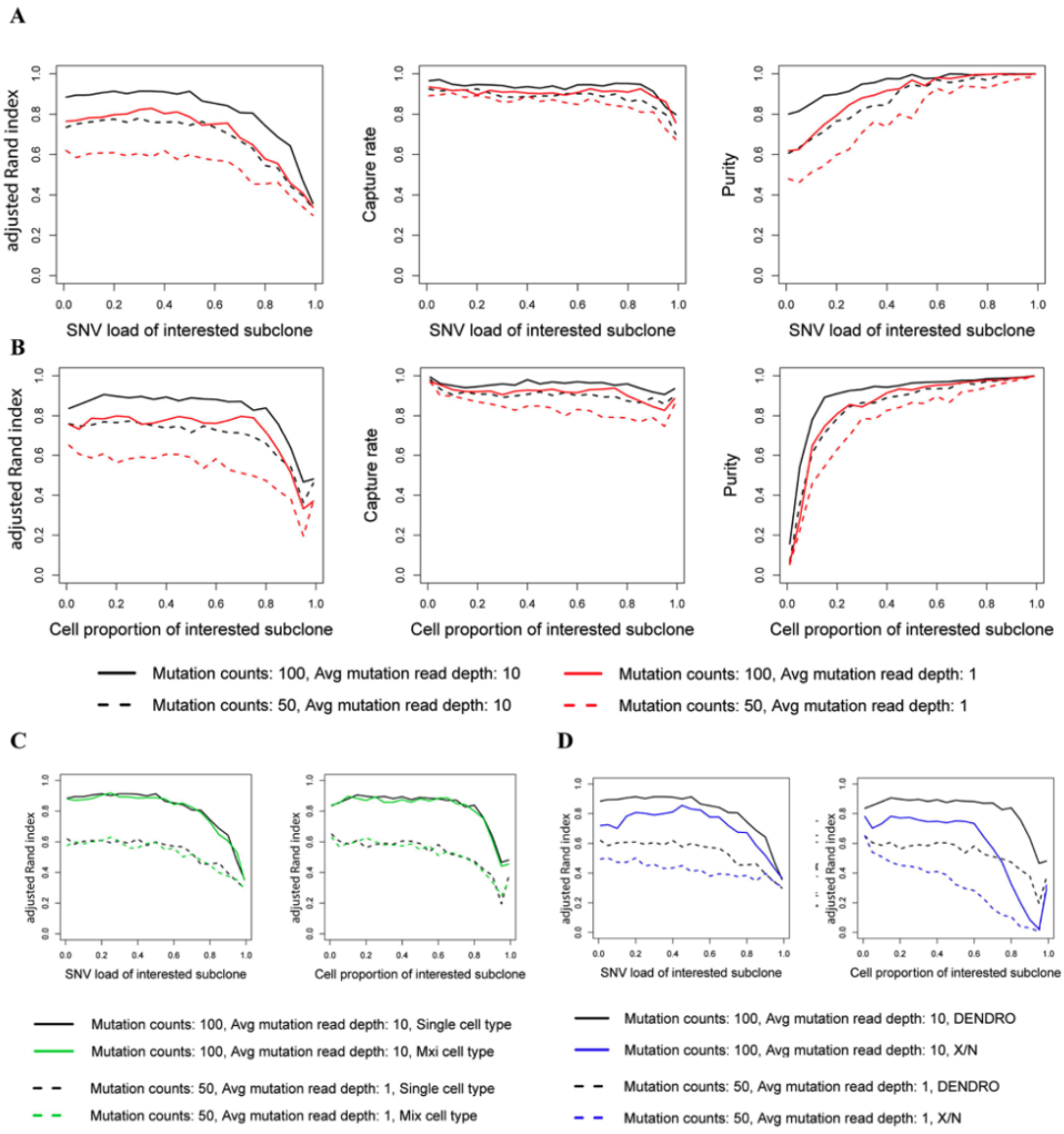


Figure 1.4 ***DENDRO accuracy assessment by simulation analysis. a*** *Statistics under different SNV load of interested clade (i.e. as fraction of total mutation counts) in a pure cell population vs. a mixture of two cell types (50% each).* ***b*** *Statistics under different cell proportion of interested clade in a pure cell population vs. a mixture of two cell types (50% each). (Mutation counts: number of mutation identified; Avg mutation read depth: average read depth of all the mutation sites.) Both plots show that mixture of cell population does not affect accuracy.* ***c*** *DENDRO accuracy assessment in mixture cell types.* ***d*** *Clustering accuracy assessment using DENDRO vs. hclust on variants allele frequencies.*

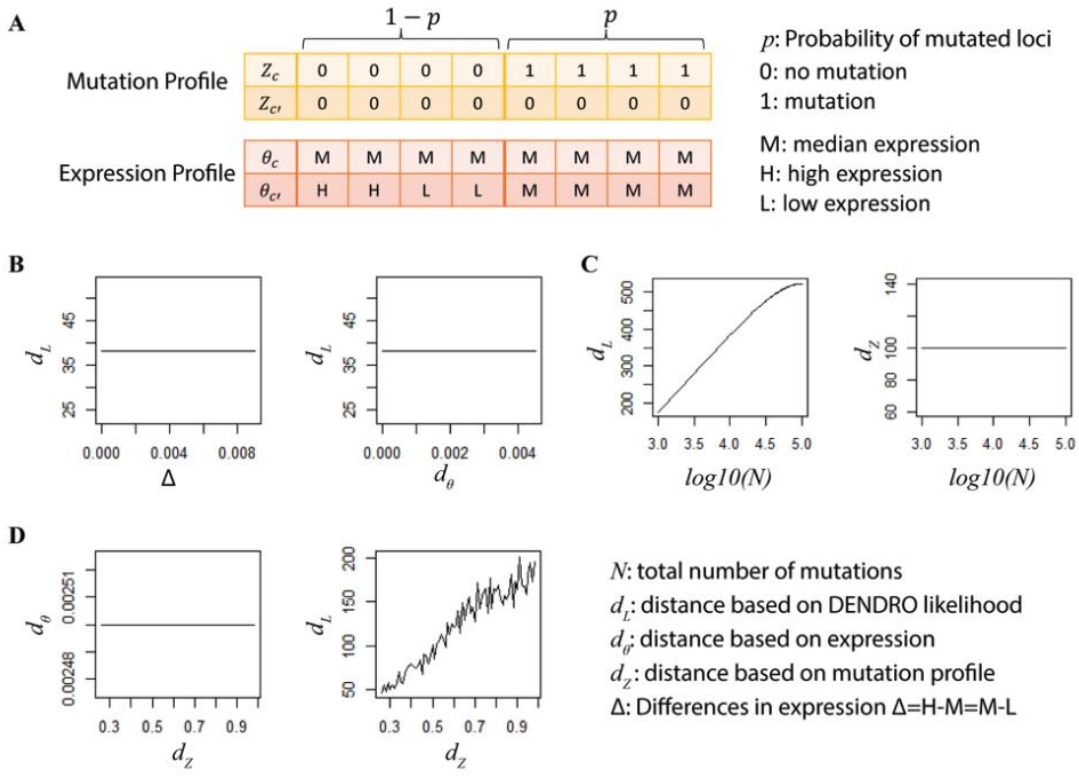


Figure 1.5 ***Kernel function justification by simulation. a*** *Illustration of simulation set up.* ***b*** *With fixed* p *and* N*, as* d_θ *increases,* d_L *stay constant. Thus, likelihood kernel is orthogonal to relative expression.* ***c*** *With fixed* $p = 1$*, as* N *approach infinite,* d_L *increases monotonically. As there are higher expression, we are more confident that the two cells belong to different clusters.* ***d*** *When* d_θ *and* N *stay the same, as* d_Z *approaches 1,* d_L *increases, because true mutation number increases.*

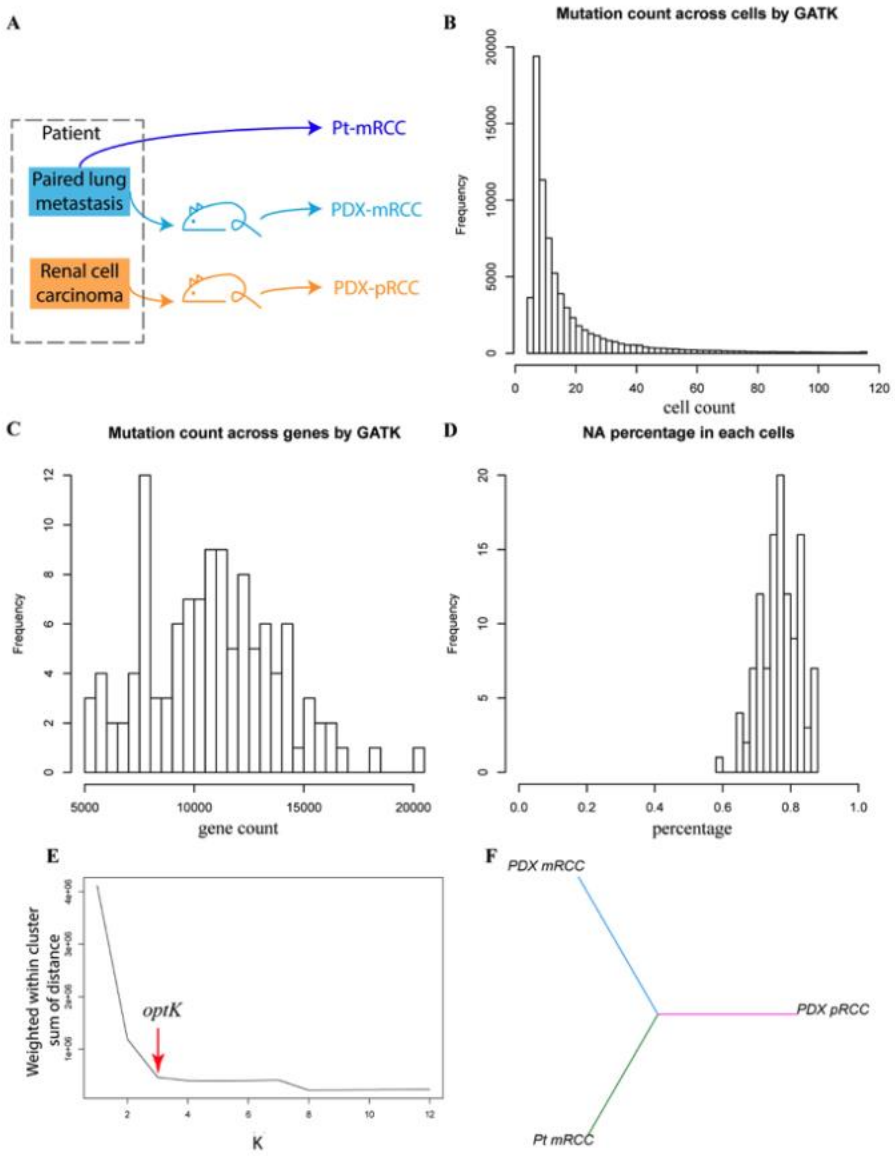


Figure 1.6 ***RCC experiment design and its mutation statistics detected by GATK tool. a*** *Experimental design for renal cell carcinoma dataset. Figure modified from Kim et al.* ***b*** *Mutation count across cells by GATK. Most of the genes have low mutation frequency.* ***c*** *Mutation count across genes by GATK tool. It shows mutation counts with a bell shape.* ***d*** *NA percentage in each cell across genes. When there is no read counts, it shows as NA.* ***e*** *Intra-cluster divergence curve to select optimal number of cluster by DENDRO. Here, optK =3.* ***f*** *Phylogenetic tree of three cell populations identified by DENDRO.*

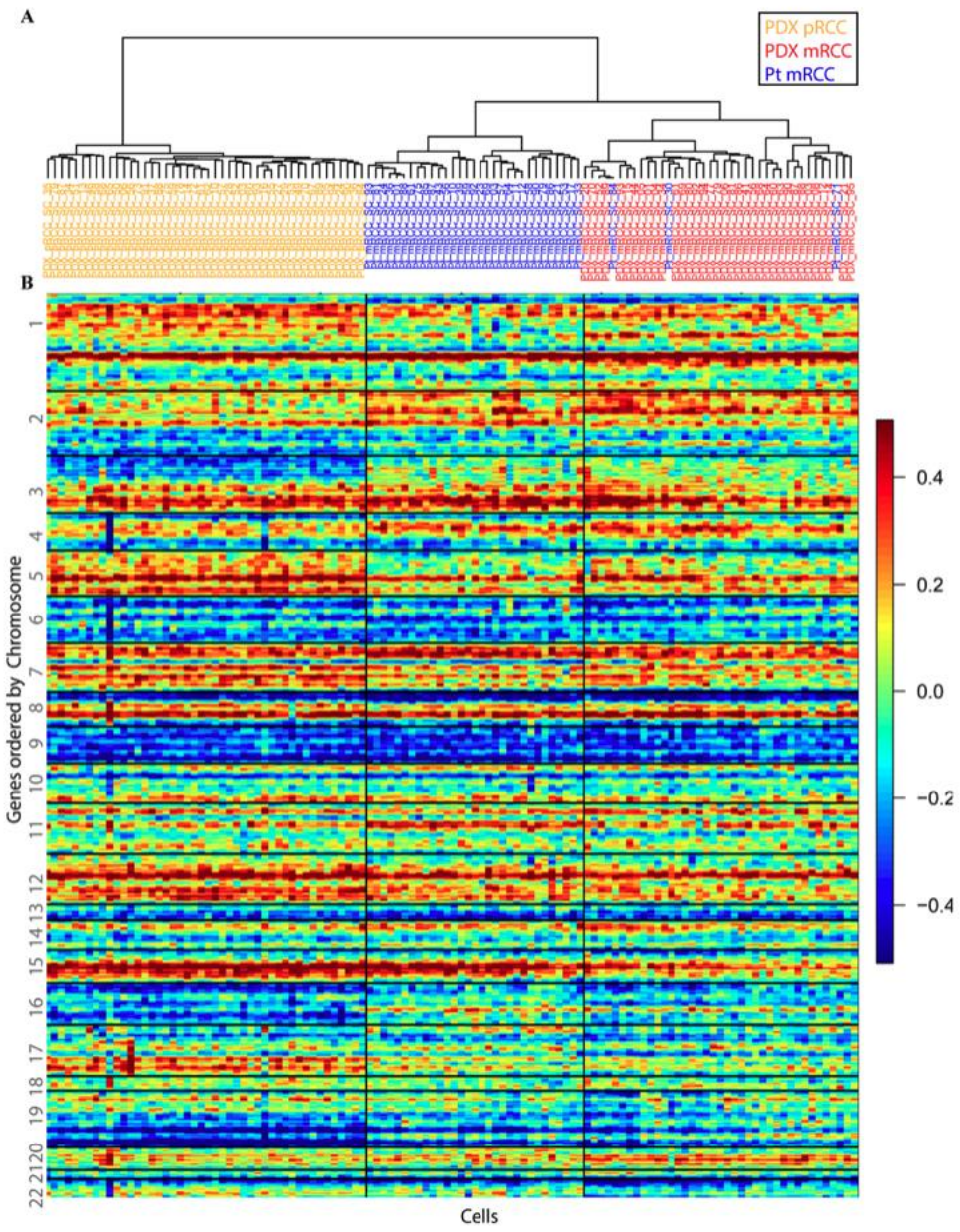


Figure 1.7 ***Expression of renal cell carcinoma. a*** *DENDRO clustering of RCC.* ***b*** *Smoothed expression ordered by DENDRO clustering. Vertical line separate cluster identified by DENDRO. Horizontal line separate different chromosomes.*

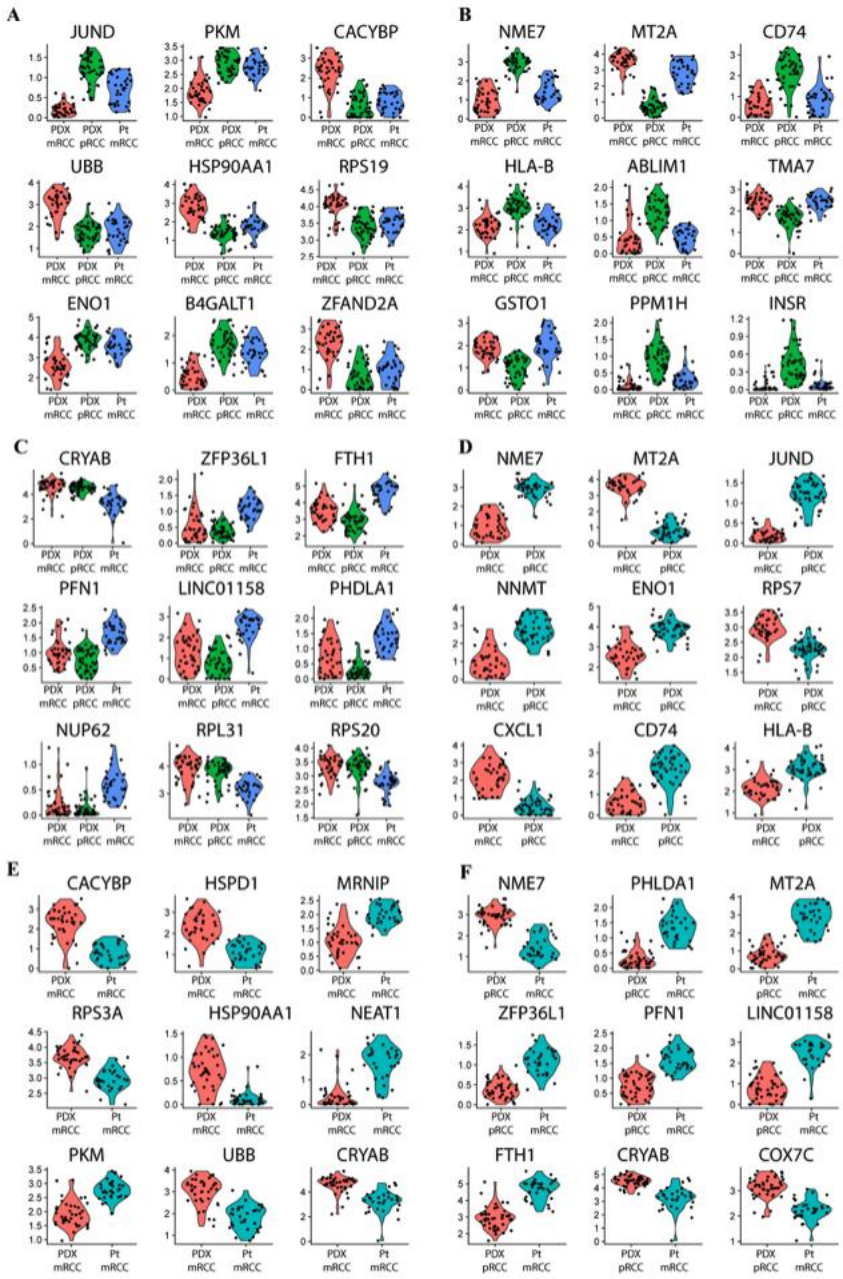


Figure 1.8 ***Most significant differential expressed genes between RCC pairs. a*** *PDX_mRCC vs. others.* ***b*** *PDX_pRCC vs. others.* ***c*** *Pt_mRCC vs. others.* ***d*** *PDX_mRCC vs. PDX_pRCC.* ***e*** *PDX_mRCC vs. Pt_mRCC.* ***f*** *Pt_mRCC vs. PDX_pRCC.*

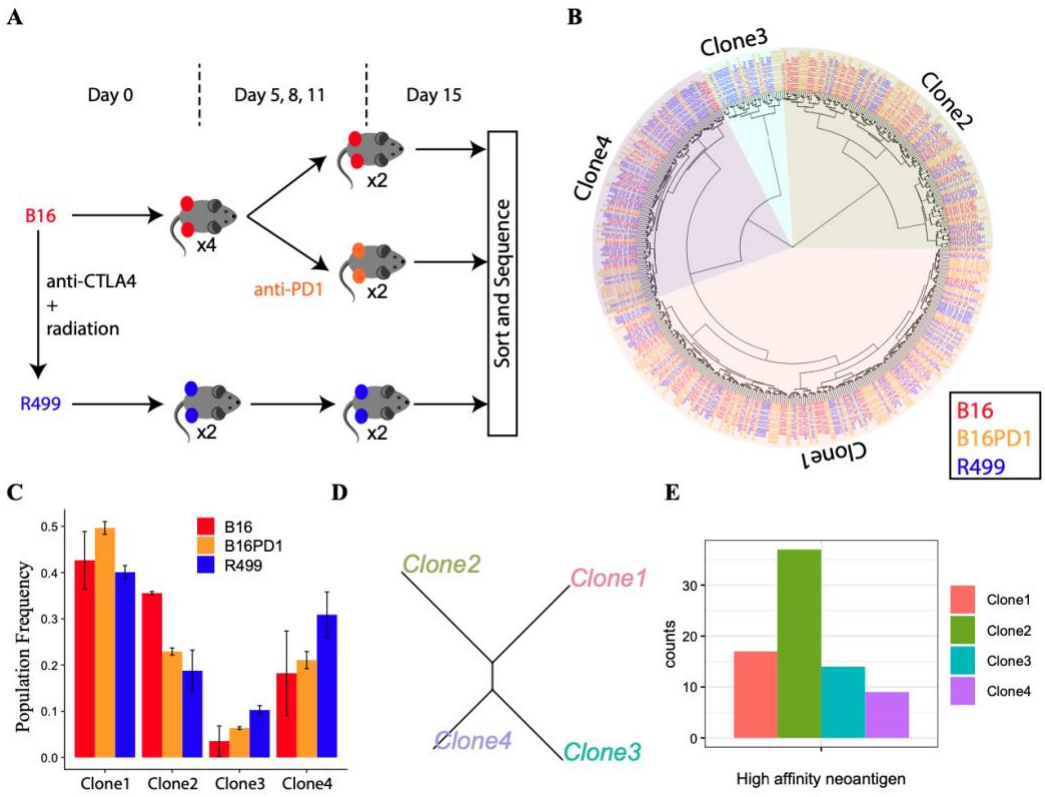


Figure 1.9 ***Clonal composition alternations with anti-PD1 treatments and cell lines. a*** *Experimental overview. For each condition at Day 15, we have two biological replicates. There are total 600 cells from 6 tumors sequenced.* ***b*** *DENDRO cluster result. No clone is exclusively associated with any tumor condition.* ***c*** *Frequencies of the subclonal populations in B16, B16PD1 and R499.* ***d*** *Neighbor joining phylogenetic tree given detected subclones.* ***e*** *Number of high affinity neoantigens predicted for each clone. Clone 2 have the highest number of neoantigens.*

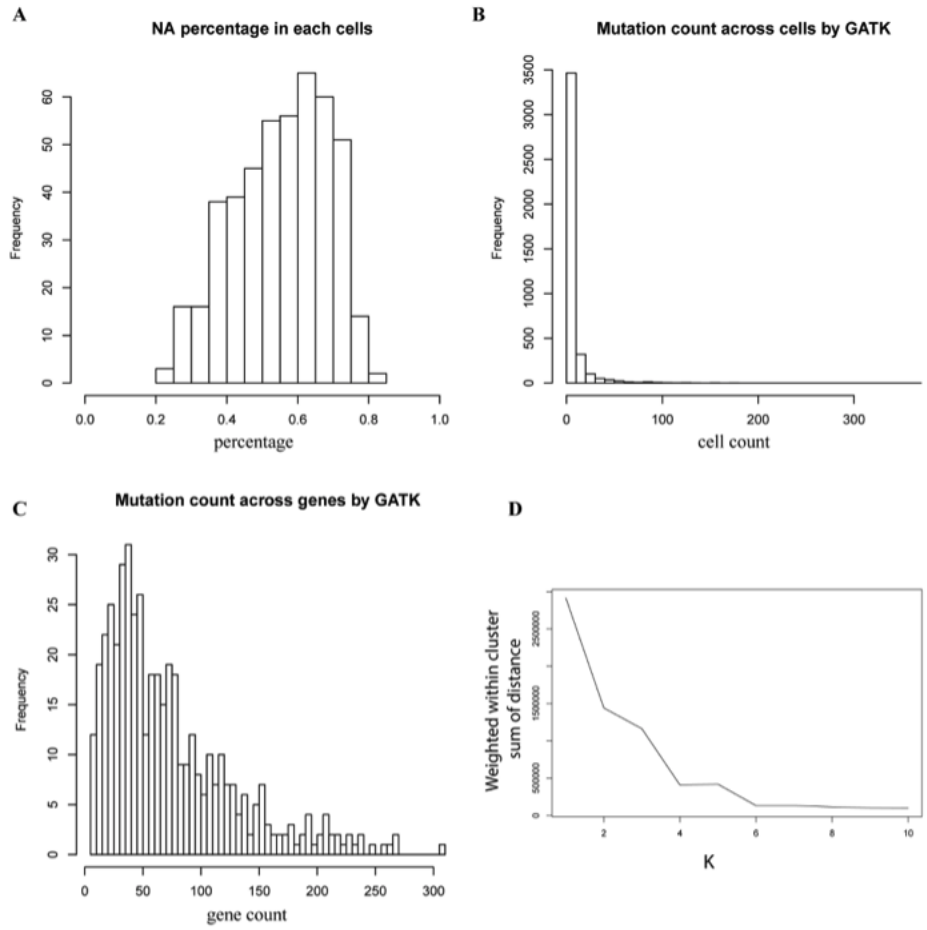


Figure 1.10 ***Anti-PD1 treatment experiment mutation statistics detected by GATK tool and optimal clustering option. a*** *NA percentage in each cell across genes. When there is no read counts, it shows as NA.* ***b*** *Mutation count across cells by GATK. Most of the genes have low mutation frequency.* ***c*** *Mutation count across genes by GATK tool. It shows mutation counts with a bell shape.* ***d*** *Intra-cluster divergence curve given different number of cluster (K). 4 is the elbow point.*

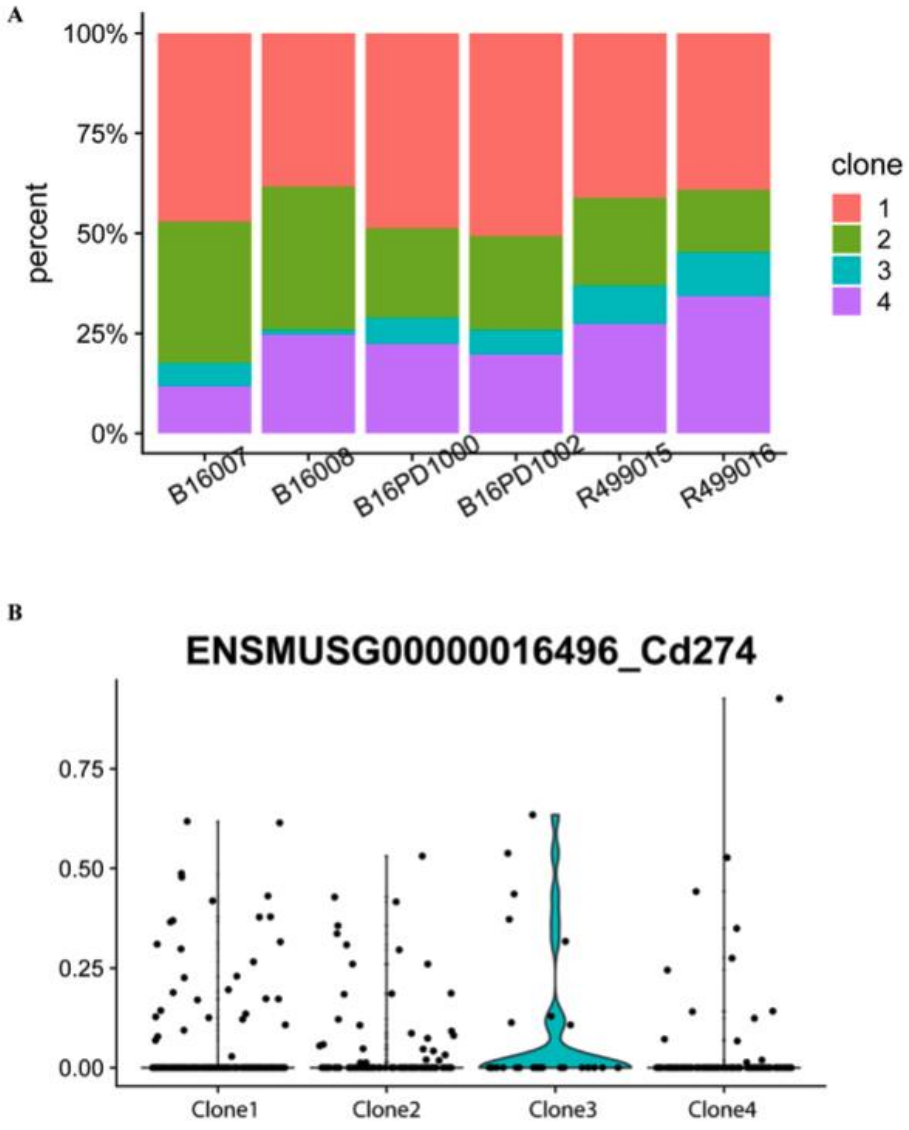


Figure 1.11 ***Anti-PD1 treatment experiment. a*** *Frequencies of the subclonal population in each of the 6 tumor samples.* ***b*** *Expression level of Pdl1 (Cd274) in each clone.*

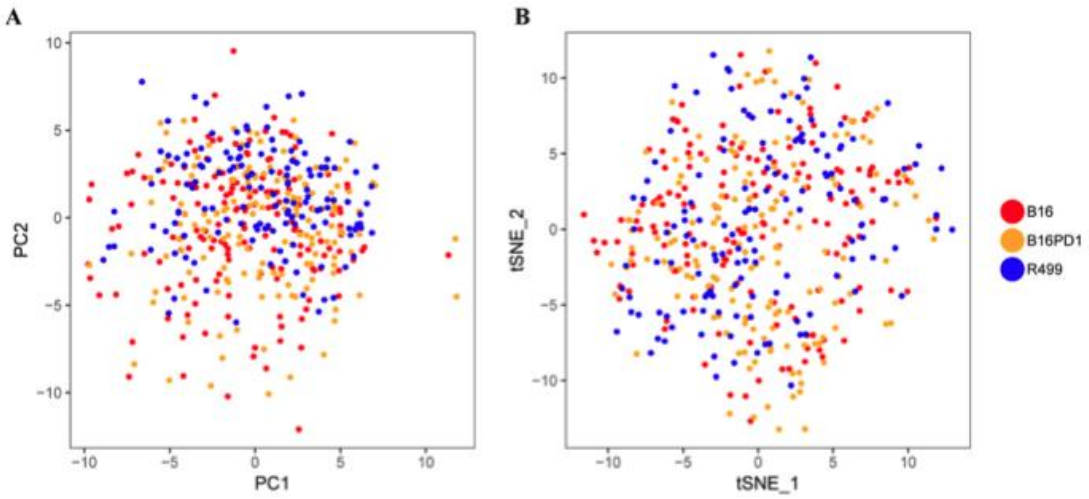


Figure 1.12 ***Transcriptome analysis on anti-PD1 treatment experiment. a*** *PCA plot of the cells based on expression.* ***b*** *t-SNE plot of the cells based on expression. Color indicates treatment regime.*

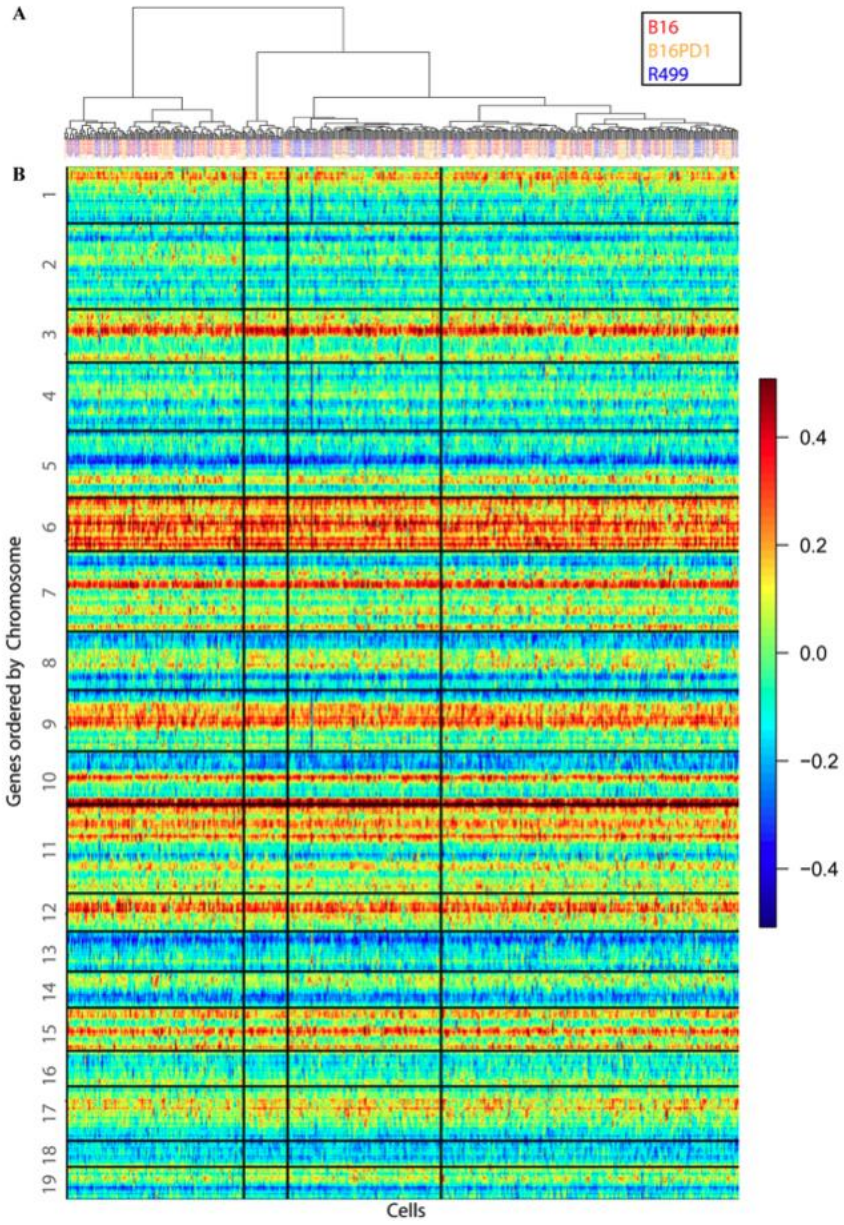


Figure 1.13 ***Expression of anti-PD1 treatment experiment. a*** *DENDRO clustering. Color indicates various conditions.* ***b*** *Smoothed expression heatmap ordered by DENDRO clustering. Vertical line separate cluster identified by DENDRO. Horizontal line separate different chromosomes.*

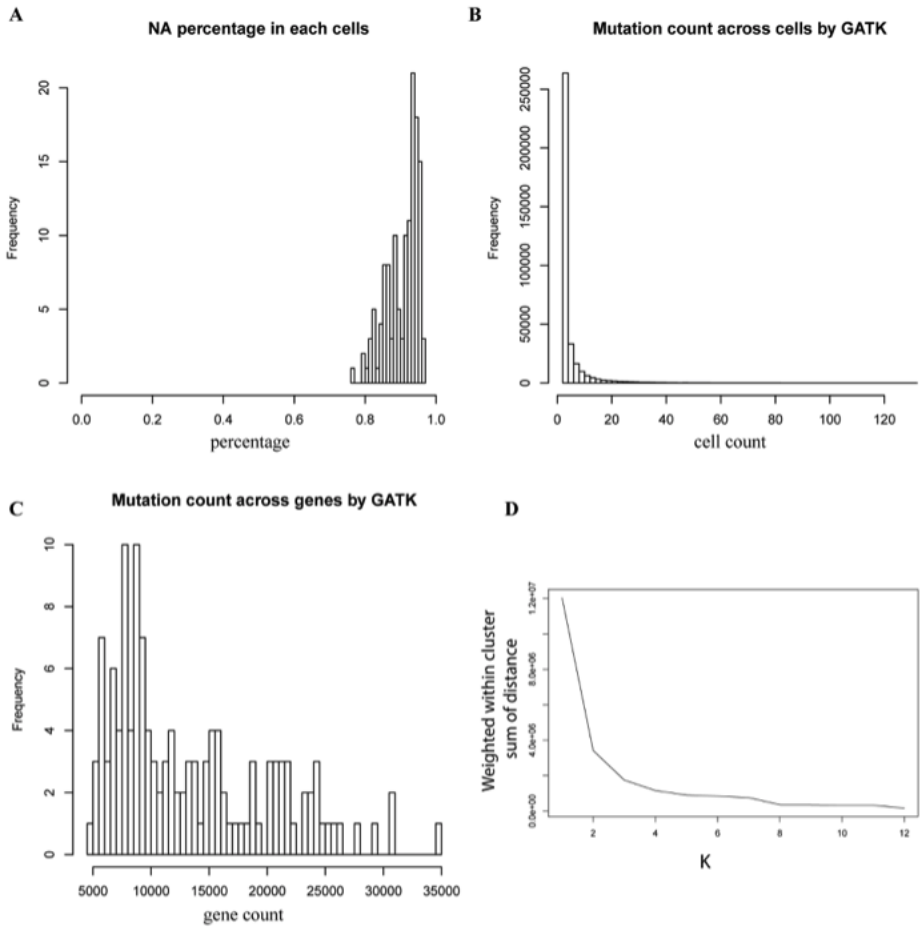


Figure 1.14 ***Breast cancer dataset mutation statistics detected by GATK tool and optimal clustering option. a*** *NA percentage in each cell across genes. When there is no read counts, it shows as NA.* ***b*** *Mutation count across cells by GATK. Most of the genes have low mutation frequency.* ***c*** *Mutation count across genes by GATK tool. It shows mutation counts with a bell shape.* ***d*** *Intra-cluster divergence curve given different number of clusters (K). 5 is the elbow point.*

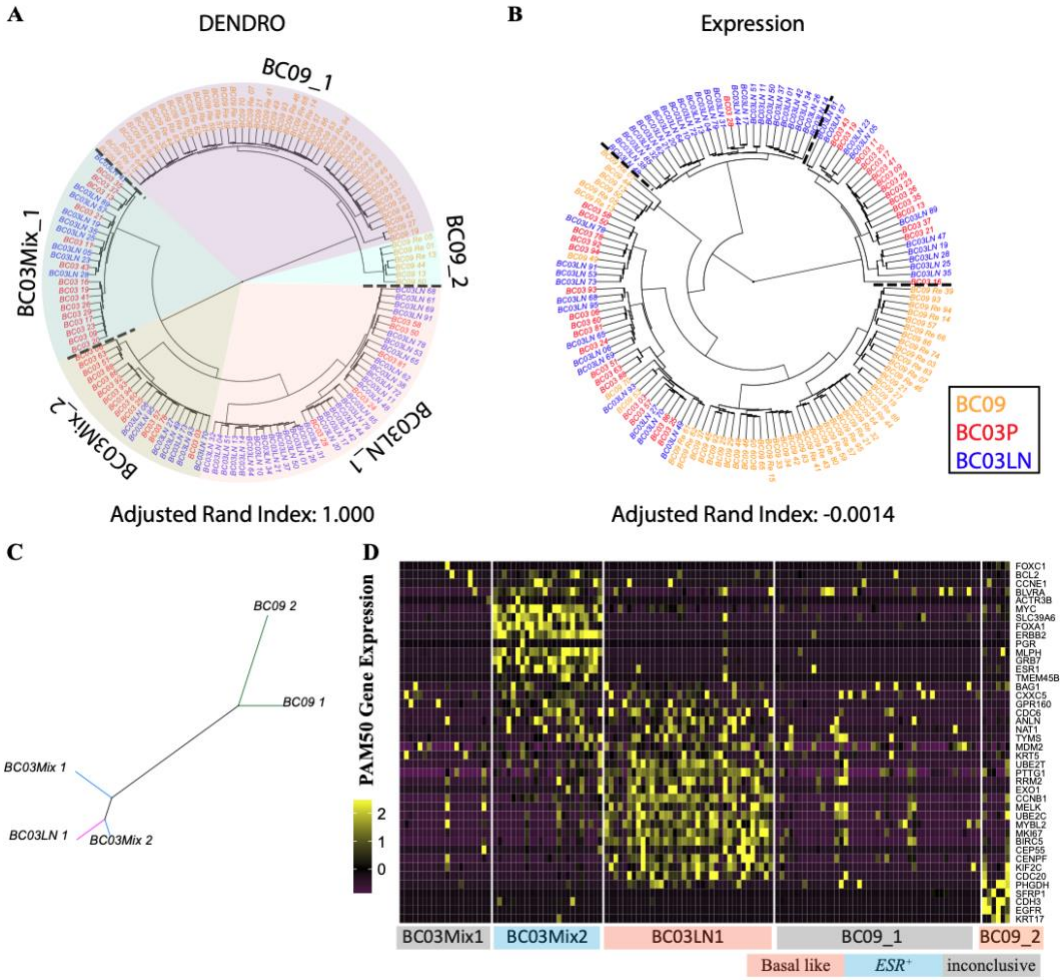


Figure 1.15 ***Analysis of scRNA-seq dataset of primary breast cancer. a*** *DENDRO cluster result for primary breast cancer dataset (Chung et al., 2017).* ***b*** *Hierarchical clustering result for the same dataset based on expression (logTPM). (dashlines indicate cluster boundaries).* ***c*** *Neighbor joining phylogenetic tree given detected subclones for breast cancer dataset.* ***d*** *PAM50 gene panel expression shows breast cancer subtypes of each subclone.*

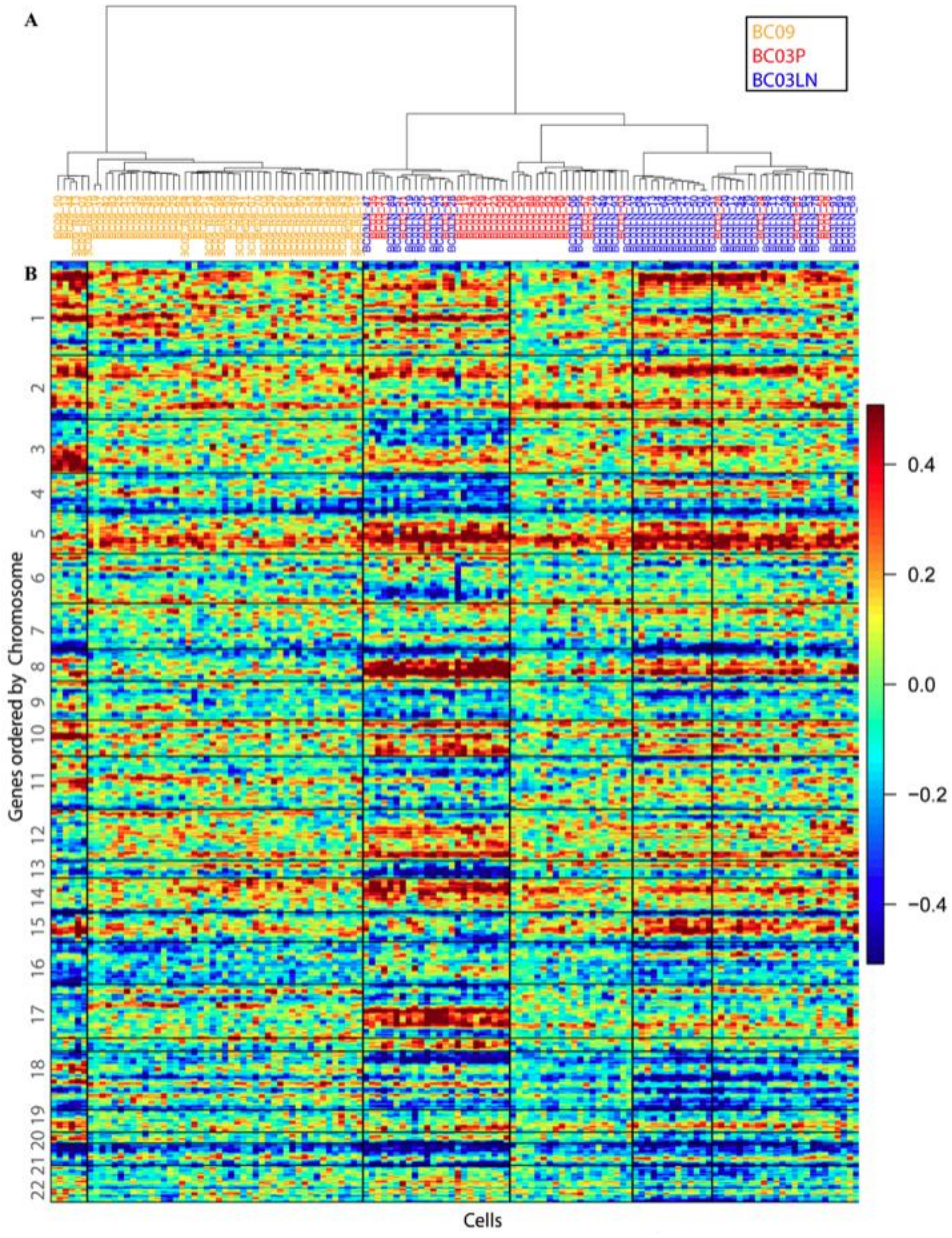


Figure 1.16 ***Expression of primary breast cancer. a*** *DENDRO clustering of breast cancer.* ***b*** *Smoothed expression ordered by DENDRO clustering. Vertical line separate cluster identified by DENDRO. Horizontal line separate different chromosomes.*

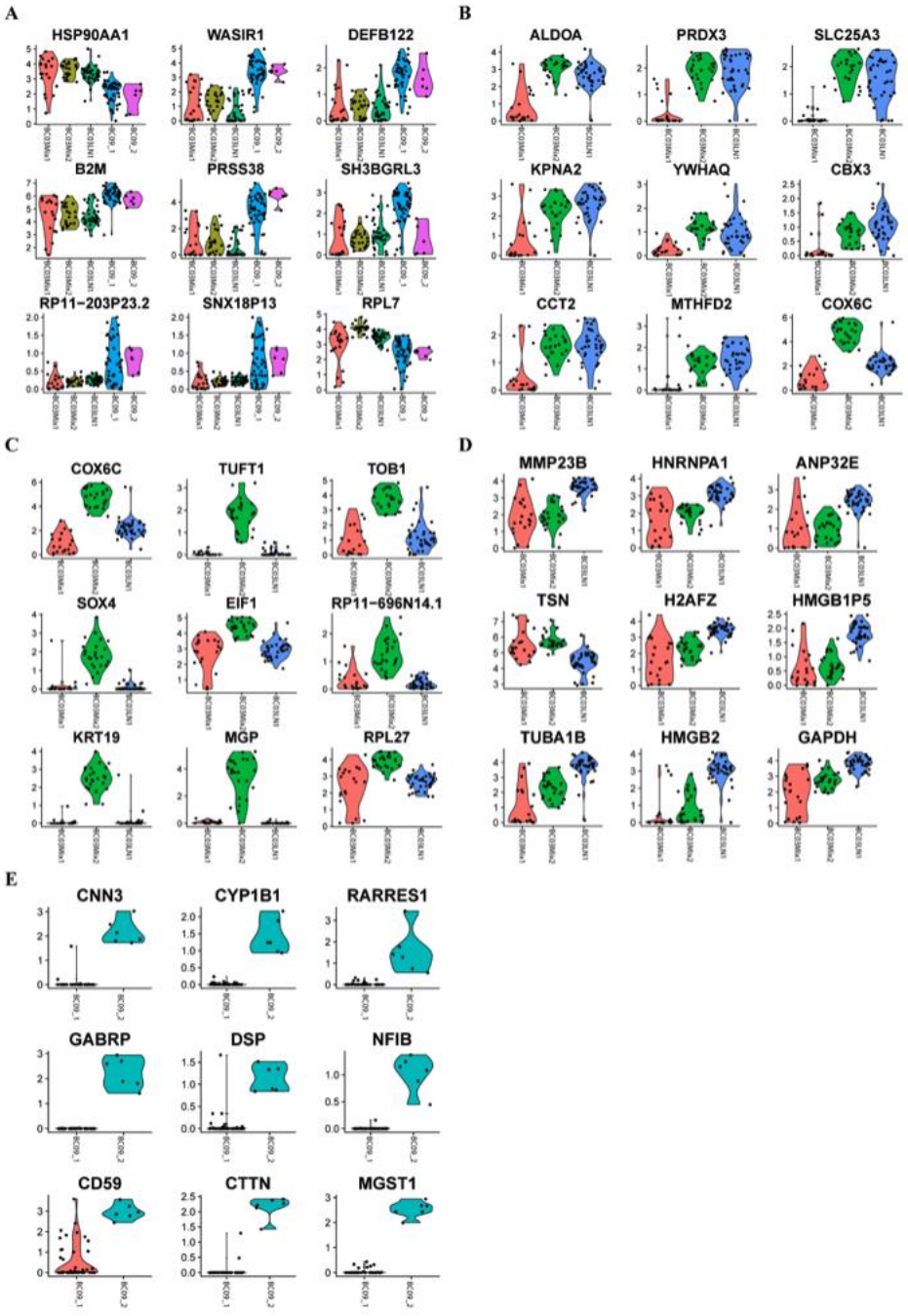


Figure 1.17 ***Most significant differential expressed genes between different BC pairs. a*** *BC03 vs. BC09.* ***b*** *BC03Mix1 vs. BC03 others.* ***c*** *BC03Mix2 vs. BC03 others.* ***d*** *BC03LN1 vs. BC03 others.* ***e*** *BC09_1 vs. BC09_2.*

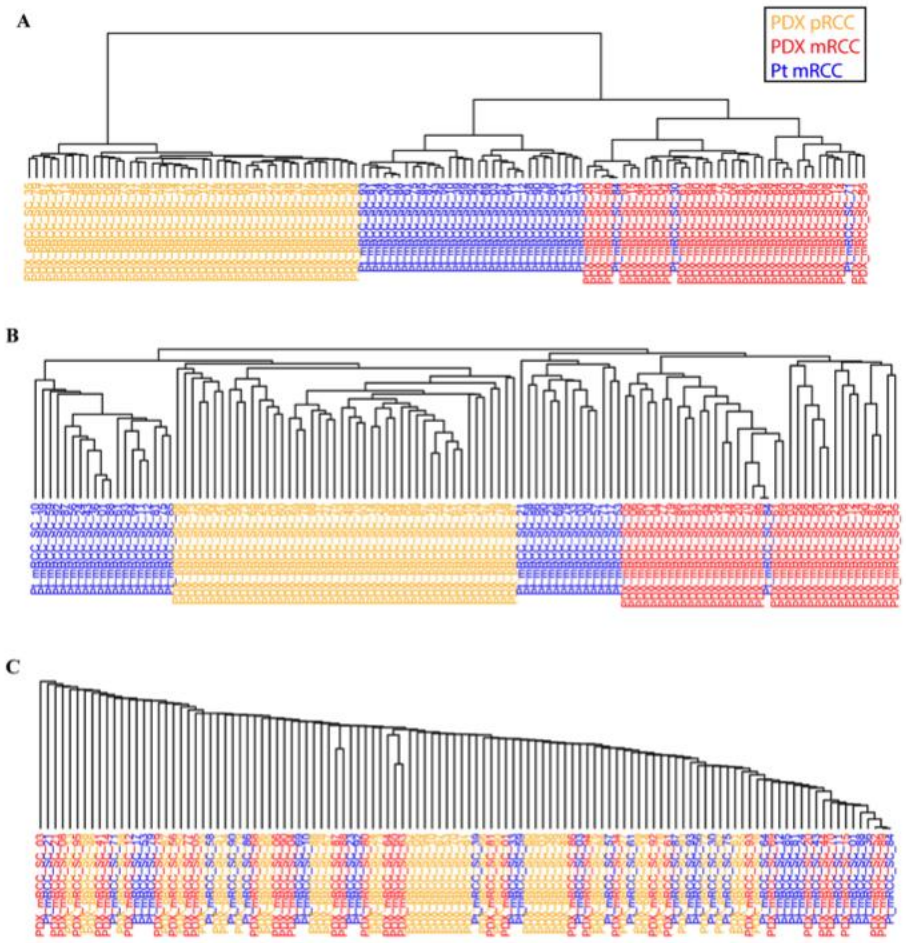


Figure 1.18 ***Hierarchical clustering algorithm comparison for renal cell carcinoma dataset.*** *Genetic divergence matrix clustering by* ***a*** *Ward.D algorithm.* ***b*** *Complete algorithm.* ***c*** *Single algorithm.*

Table 1.1 ***a*** *RCC subclone cell composition and labels.* ***b*** *BC subclone cell composition and labels.*

a	**Cluster 1**	**Cluster 2**	**Cluster 3**
PDX_mRCC	36	0	0
PDX_pRCC	0	46	0
Pt_mRCC	3	0	31
Final label	PDX_mRCC	PDX_pRCC	Pt_mRCC

b	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**
BC03	13	15	5	0	0
BC03LN	7	9	32	0	0
BC09	0	0	0	45	6
Final label	BC03Mix_1	BC03Mix_2	BC03LN_1	BC09_1	BC09_2

Table 1.2 ***a*** *Number of differential expressed gene between groups.* ***b*** *Number of differential expressed gene between groups overlapped with differential mutated genes (# of overlapped genes/# of differential expressed genes).*

a				
	PDX_mRCC	PDX_pRCC	Pt_mRCC	Other
PDX_mRCC		276	74	181
PDX_pRCC	276		191	302
Pt_mRCC	74	191		98
Other	181	302	98	
b				
	PDX_mRCC	PDX_pRCC	Pt_mRCC	Other
PDX_mRCC		93/276	24/74	41/181
PDX_pRCC	93/276		68/191	64/302
Pt_mRCC	24/74	68/191		15/98
Other	41/181	64/302	15/98	

Table 1.3 ***GO analysis on Differential Expressed Genes between Pt_mRCC and PDX_mRCC***

Gene Set Name [# Genes (K)]				**p-value**	**FDR q-value**
Color key:	Cancer-related pathway	Immune-related pathway	Other		
HALLMARK_TNFA_SIGNALING_VIA_NFKB [200]				1.86 e-8	9.28 e-7
HALLMARK_HYPOXIA [200]*				5.04 e-7	1.26 e-5
HALLMARK_MTORC1_SIGNALING [200]*				1.15 e-5	1.92 e-4
HALLMARK_COMPLEMENT [200]				2.15 e-4	1.79 e-3
HALLMARK_GLYCOLYSIS [200]*				2.15 e-4	1.79 e-3
HALLMARK_KRAS_SIGNALING_UP [200]				2.15 e-4	1.79 e-3
HALLMARK_ALLOGRAFT_REJECTION [200]				3.17 e-3	1.58 e-2
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSINSITION [200]*				3.17 e-3	1.58 e-2
HALLMARK_ESTROGEN_RESPONSE_EARLY [200]				3.17 e-3	1.58 e-2
HALLMARK_INFLAMMATORY_RESPONSE [200]				3.17 e-3	1.58 e-2

Table 1.4 ***GO analysis on Differential Mutated Genes between Pt_mRCC and PDX_mRCC***

Gene Set Name [# Genes (K)]				**p-value**	**FDR q-value**
Color key:	Cancer-related pathway	Immune-related pathway	Other		
HALLMARK_UV_RESPONSE_DN [144]				1.39 e-29	6.93 e-28
HALLMARK_MYC_TARGETS_V1 [200]				2.75 e-27	6.87 e-26
HALLMARK_MITOTIC_SPINDLE [200]				7.97 e-24	1.33 e-22
HALLMARK_MTORC1_SIGNALING [200]*				9.49 e-20	1.19 e-18
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSINSITION [200]*				1.91 e-17	1.91 e-16
HALLMARK_OXIDATIVE_PHOSPHORYLATION [200]				1.06 e-16	8.81 e-16
HALLMARK_HYPOXIA [200]*				5.69 e-16	4.06 e-15
HALLMARK_GLYCOLYSIS [200]*				2.97 e-15	1.86 e-14
HALLMARK_ANDROGEN_RESPONSE [101]				1.07 e-14	5.96 e-14
HALLMARK_HEME_METABOLISM [200]				7.4 e-14	3.7 e-13

Table 1.5 ***Mean expression correlation between samples: Chung et al. 2017***

	BC03	BC09	BC03LN
BC03		**0.595**	**0.826**
BC09			**0.648**
BC03LN			

Table 1.6 ***a*** *Number of differential expressed gene between groups.* ***b*** *Number of differential expressed gene between groups overlapped with differential mutated genes*

a	BC03	BC09	BC03 Mix1	BC03 Mix2	BC03 LN1	BC0 9_1	BC0 9_2	Other within same tumor
BC03		178						
BC09	178							
BC03Mix1								110
BC03Mix2								322
BC03LN1								183
BC09_1							34	
BC09_2						34		
Other within same tumor			110	322	183			

b	BC03	BC09	BC03 Mix1	BC03 Mix2	BC03 LN1	BC0 9_1	BC0 9_2	Other within same tumor
BC03		102/ 178						
BC09	102/ 178							
BC03Mix1								71/110
BC03Mix2								206/322
BC03LN1								111/183
BC09_1							21/3 4	
BC09_2						21/3 4		
Other within same tumor			71/11 0	206/3 22	111/1 83			

CHAPTER 2 SURFACE PROTEIN IMPUTATION FROM SINGLE CELL TRANSCRIPTOMES BY DEEP NEURAL NETWORK

2.1 Introduction

Recent technological advances allow the simultaneous profiling, across many cells in parallel, of multiple omics features in the same cell [29, 87-90]. In particular, high throughput quantification of the transcriptome and a selected panel of cell surface proteins in the same cell is now feasible through the REAP-seq and CITE-seq protocols [88, 89]. Cell surface proteins can serve as integral markers of specific cellular functions and primary targets for therapeutic intervention. Immunophenotyping by cell surface proteins has been an indispensable tool in hematopoiesis, immunology and cancer research during the past 30 years. Yet, due to technological barriers and cost considerations, most single cell studies, including Human Cell Atlas project [91], quantify the transcriptome only and do not have cell-matched measurements of relevant surface proteins [22, 92]. Sometimes, which cell types and corresponding surface proteins are essential become apparent only after exploration by scRNA-seq. This motivates our inquiry of whether protein abundances in individual cells can be accurately imputed by the cell's transcriptome.

We propose cTP-net (single cell Transcriptome to Protein prediction with deep neural network), a transfer learning approach based on deep neural networks that imputes surface protein abundances for scRNA-seq data. Through comprehensive benchmark evaluations and applications to Human Cell Atlas and acute myeloid leukemia data sets, we show that cTP-net outperform existing methods and can transfer information from training data to accurately impute 24 immunophenotype markers, which achieve a more detailed characterization of cellular state and cellular phenotypes than transcriptome measurements alone. cTP-net relies, for model training, on accumulating public data of cells with paired transcriptome and surface protein measurements.

2.2 Results

2.2.1 Method overview

An overview of cTP-net is shown in Figure 2.1a. Studies based on both CITE-seq and REAP-seq have shown that the relative abundance of most surface proteins, at the level of individual cells, is only weakly correlated with the relative abundance of the RNA of its corresponding gene [88, 89, 93]. This is due to technical factors such as RNA and protein measurement error [94], as well as inherent stochasticity in RNA processing, translation and protein transport [95-99]. To accurately impute surface protein abundance from scRNA-seq data, cTP-net employs two steps: (1) denoising of the scRNA-seq count matrix and (2) imputation based on the denoised data through a transcriptome-protein mapping (Figure 2.1a). The initial denoising, by SAVER-X [100], produces more accurate estimates of the RNA transcript relative abundances for each cell. Compared to the raw counts, the denoised relative expression values have significantly improved correlation with their corresponding protein measurement (Figure 2.1b, 2.2a, 2.3ab). Yet, for some surface proteins, such as CD45RA, this correlation for denoised expression is still extremely low.

The production of a surface protein from its corresponding RNA transcript is a complicated process involving post-transcriptional modifications and transport [95], translation [96], post-translational modifications [97] and protein trafficking [98]. These processes depend on the state of the cell and the activities of other genes [93, 99]. To learn the mapping from a cell's transcriptome to the relative abundance of a given set of surface proteins, cTP-net employs a multiple branch deep neural network (MB-DNN, Figure 2.4). Deep neural networks have recently shown success in modeling complex biological systems [101, 102], and more importantly, allow good generalization across data sets[100, 103]. Generalization performance is an important aspect of cTP-net, as we would like to perform imputation on tissues that do not exactly match

the training data in cell type composition. Details of the cTP-net model and training procedure, as well as of alternative models and procedures that we have tried, are in Methods.

2.2.2 Imputation accuracy evaluation via random holdout

To examine imputation accuracy, we first consider the ideal case where imputation is conducted on cells of types that exactly match those in training data. For benchmarking, we used peripheral blood mononuclear cells (PBMCs) and cord blood mononuclear cells (CBMCs) processed by CITE-seq and REAP-seq [88, 89], described in Table 2.1. We employed holdout method, where the cells in each data set were randomly partitioned into two sets: a training set with 90% of the cells and a holdout set with the remaining 10% of the cells for validation (Methods, Figure 2.5a). Each cell type is well represented in both the training and validation sets. Figure 2.1b and S3a show that, for all proteins examined in the CITE-seq PBMC data, cTP-net imputed abundances have much higher correlation to the measured protein levels, as compared with the denoised and raw RNA counts of the corresponding genes. We obtained similar results for the CITE-seq CBMC and REAP-seq PBMC data sets (Figure 2.3ab).

2.2.3 Generalization accuracy to unseen cell types

Next, we considered the generalization accuracy of cTP-net, testing whether it produces accurate imputations for cell types that are not present in the training set. For each of the high-level cell types in each data set in Table 2.2, all cells of the given type are held out during training, and cTP-net, trained on the rest of the cells, was then used to impute protein abundances for the held out cells (Methods, Figure 2.5b). We did this for each cell type and generated an “out-of-cell-type” prediction for every cell.

Across all benchmarking data sets and all cell types, these out-of-cell-type predictions still improve significantly upon the corresponding RNA counts while slightly inferior in accuracy to the traditional holdout validation predictions above (Figure 2.6a, 2.3a). This indicates that cTP-net

provides informative predictions on cell types not present during training, vastly improving upon using the corresponding mRNA transcript abundance as proxy for the protein level.

2.2.4 Generalization accuracy across tissue and lab protocol

To further examine the case where cell types in the training and test data are not perfectly aligned, we considered a scenario where the model is applied to perform imputation on a tissue that differs from the training data. We trained cTP-net on PBMCs and then applied it to perform imputation on CBMCs, and vice versa, using the data from Stoeckius et al. [89] (Methods). Cord blood is expected to be enriched for stem cells and cells undergoing differentiation, whereas peripheral blood contains well-differentiated cell types, and thus the two populations are composed of different but related cell types. Figure 2.6a and 2.2b shows the result on training on CBMC and then imputing on PBMC. Imputing across tissue markedly improves the correlation to the measured protein level, as compared to the denoised RNA of the corresponding gene, but is worse than imputation produced by model trained on the same population. For practical use, we have trained a network using all cell populations combined, which indeed achieves better accuracy than a network trained on each separately (Methods, Figure 2.2b, 2.3ac). The weights for this network are publicly available at https://github.com/zhouzilu/cTPnet.

We then tested whether cTP-net's predictions are sensitive to the laboratory protocol, and in particular, whether networks trained using CITE-seq data yields good predictions by REAP-seq's standard, and vice versa. Using a benchmarking design similar to above, we found that, in general, cTP-net maintains good generalization power across these two protocols (Figure 2.6a, 2.2b).

2.2.5 Imputation accuracy comparison to Seurat v3

Seurat v3 anchor transfer [104] is a recent approach that uses cell alignment between data sets to impute features for single cell data. For comparison, we applied Seurat v3 anchor transfer to the holdout validation and out-of-cell-type benchmarking scenarios above (Methods). In the

validation scenario, we found the performance of cTP-net and Seurat v3 to be comparable, with cTP-net slightly better, as both methods can estimate protein abundance by utilizing marker genes to identify the cell types. cTP-net, however, vastly improves upon Seurat in the out-of-cell-type scenario (Figure 2.6a, 2.7a). This is because cTP-net's neural network, trained across a diversity of cell types, learns a direct transcriptome-protein mapping that can more flexibly generalize to unseen cell types, while Seurat v3 depends on a nearest neighbor method that can only sample from the training dataset. As shown by the cross-population and out-of-cell-type benchmarking above, cTP-net does not require direct congruence of cell types across training and test sets.

In addition to predictions on unseen cell type, cTP-net also improves upon the existing state-of-the-art in capturing within cell-type variation in protein abundance. As expected, within cell-type variation is harder to predict, but cTP-net's imputations nevertheless achieve high correlations with measured protein abundance for a subset of proteins and cell types (Figure 2.2c, 2.3d). Compared to Seurat v3, cTP-net's imputations align more accurately with measured protein levels when zoomed into cells of the same type (Figure 2.6b, 2.7b); see Figure 2.6c, for example, CD11c in CD14-CD16+ monocytes, CD2 in CD8 T cells, and CD16 in dendritic cells. All of these surface proteins have important biological function in the corresponding cell types, as CD11c helps trigger respiratory burst in monocyte [105], CD2 co-stimulates molecule on T cells [106] and CD16 differentiate DC subpopulation [107]. The learning of such within-type heterogeneity gives cTP-net the potential to attain higher resolution in the discovery and labeling of cell states.

2.2.6 Network interpretation and feature importance

What types of features are being used by cTP-net to form its imputation? To interpret the network, we conducted a permutation-based interpolation analysis, which calculates a permutation feature importance for each protein-gene pair (Methods, Figure 2.8a). Interpolation

can be done using all cells, or cells of a specific type, the latter allowing us to probe relationships that may be specific to a given cell type. Applying this analysis to cTP-net trained on PBMC, we found that, at the level of the general population that includes all cell types, the most important genes for the prediction of each protein are those that exhibit the highest cell-type specificity in expression (Table 2.3). This is because most of these surface proteins are cell type markers, and thus when cells of all types are pooled together, “cell type” is the key latent variable that underlies their heterogeneity. In addition, as cell-type markers are usually redundant and predictable by other genes, the model still performs well after removing corresponding surface protein genes during training (Table 2.4, 2.5). Within cell type interpolation, on the other hand, reveals genes related to RNA processing, RNA binding, protein localization and biosynthetic processes, in addition to immune-related genes that differentiate the immune cell sub-types (Table 2.6). This analysis shows that cTP-net combines different types of features, both cell type markers and genes involved in RNA to protein conversion and transport, to achieve multiscale imputation accuracy.

In addition, we analyzed the bottleneck layer with 128 nodes before the network branched out to the protein-specific layers. We performed dimension reduction (UMAP) directly on the bottleneck layer intermediate output of 7000 PBMCs from CITE-seq. Figure 2.8b shows that the cells are cleanly separated into different clusters, representing cell types as well as gradients in surface protein abundance. This confirms that the bottleneck layer captures the essential information on cell stages and transitions, and that each subsequent individual branch then predicts its corresponding protein’s abundance.

2.2.7 Application to Human Cell Atlas

Having benchmarked cTP-net’s generalization accuracy across immune cell types, tissues, and technologies, we then applied the network trained on the combined CITE-seq dataset of PBMCs,CBMCs and bone marrow mononuclear cells (BMMCs) [89, 104] to perform imputation

for the Human Cell Atlas CBMC and BMMC data sets (Table 2.1). Figure 2.9 shows the raw RNA count and predicted surface protein abundance for 24 markers across 6023 BMMCs from sample MantonBM1 and 4176 CBMCs from sample MantonCB1. (Similar plots for the other 7 BMMC and 7 CBMC samples are shown in Figure 2.10, 2.11). Similar to what was observed for actual measured protein abundances in the CITE-seq and REAP-seq studies, the imputed protein levels differ markedly from the RNA expression of its corresponding gene, displaying higher contrast across cell types and higher uniformity within cell type. Thus, the imputed protein levels serve as interpretable intermediate features for the identification and labelling of cell states, defining cell subtypes more clearly than the RNA levels of the corresponding marker genes. For example, imputed CD4 and CD8 levels separate CD4+ T cells from CD8+ T cells with high confidence. Further separation of naïve T cells to memory T cells can be achieved through imputed CD45RA/CD45RO abundance, as CD45RA is a naïve antigen and CD45RO is a memory antigen. Consistent with flow cytometry data, the large majority of CB T cells are naïve, whereas the BM T cell population is more diverse [108]. Also, for BM B cells that have high imputed CD19 levels, cTP-net allows us to confidently distinguish the Pre.B (CD38+, CD127+), immature B (CD38+, CD79b+), memory B (CD27+) and naïve B cells (CD27-), whose immunophenotypes have been well characterized [109].

In addition, consider natural killer cells, in which the proteins CD56 and CD16 serve as indicators for immunostimulatory effector functions, including an efficient cytotoxic capacity [110, 111]. We observe an opposing gradient of imputed CD56 and CD16 levels within transcriptomically derived natural killer (NK) cell clusters that reveal $CD56_{bright}$ and $CD56_{dim}$ subsets, coherent with previous studies[89] (Figure 2.6f, 2.12, F-test: p-value = 1.667e-15). This pattern is not found in RNA abundances due to low expression (F-test: p-value= 0.9377). Between $CD56_{brignt}$ and $CD56_{dim}$ subsets, 7 out of 10 of previously studied differentially expressed genes are significant in the single cell analysis (Fisher test: p-value = 1.07e-04) [89, 112, 113].

This gradient in CD56 and CD16, where decrease in CD56 is accompanied by increase in CD16, is replicated across the 8 CBMC and 8 BMMC samples in HCA (Figure 2.10, 2.11, 2.12).

Consider also the case of CD57, which is a marker for terminally differentiated "senescent" cells in the T and NK cell types. The imputed level of CD57 is lower in CBMCs (fetus's blood), and rises in BMMCs (95% quantile: bootstrap p-value<1e-6). This is consistent with expectation since CD57+ NK cell and T cell populations grow after birth and with ageing [114-116] (Figure 2.10, 2.11).

These results demonstrate how cTP-net, trained on a combination of PBMCs, CBMCs and BMMCs, can impute cell type, cell stage, and tissue-specific protein signatures in new data without explicitly being given the tissue of origin.

2.2.8 Application to Acute Myeloid Leukemia

We further apply cTP-net to an acute myeloid leukemia (AML) data set from Galen et al. [31]. AML is a heterogeneous disease where the diversity of malignant cell types partially recapitulates the stages of myeloid development. Mapping the malignant cells in AML to the differentiation stage of their cell of origin strongly impacts tumor prognosis and treatment, as malignant cells that originate from earlier stage progenitors have higher risk of relapse [117, 118]. In the original paper, the authors sequenced 7698 cells from 5 healthy donors to build a reference map of cell types during myeloid development, and then mapped 30712 cells from 16 AML patients across multiple time points to this reference to identify the differentiation stage of the malignant cells. Here, by imputing 24 immunophenotype markers with cTP-net, we can directly characterize the differentiation stage of cell-of-origin for the malignant cells.

Figure 2.13a is a UMAP plot based on imputed surface protein abundance of 5 normal BMs and 12 Day 0 samples from AML patients. The majority of the malignant cells as identified in the original paper reside on the right half of the plot, which recapitulate the myeloid differentiation

trajectory as revealed by the imputed values of canonical protein markers (Figure 2.13b): From CD34+ progenitors to CD38+CD123+ cells in transition to CD11c+ and CD14+ mature monocytes [119]. All of the malignant cells have imputed protein values that place them along this monocyte lineage. Using the transcriptome for visualization, on the other hand, reveals large batch effects across samples, due to both technical batch and biological differences (Figure 2.14). Thus, unlike the imputed protein data, the transcriptomic data cannot be directly combined without alignment.

Based on the trajectory revealed by the imputed protein levels, we can determine the differentiation cell stage(s) for the malignant cells of each tumor, according to which the 12 AML patients can be divided into three categories: (1) AMLs of single differentiation stage (AML420B, AML556, AML707B and AML916; Figure 2.13c), (2) AMLs of two differentiation stages (AML210A, AML328, AML419A and AML475; Figure 2.13e) and (3) AMLs of many differentiation stages (AML1012, AML329, AML870 and AML921A; Figure 2.13f). This stage assignment is consistent with the original study [31]. For example, AML419A harbors two malignant cell types at opposite ends of the monocyte differentiation axis, distinguished by imputed CD34 and CD11c levels as CD34+CD11c- indicates progenitor-like and CD34-CD11c+ indicates differentiated monocyte-like cells (Figure 2.13d, 4e). AML707B, which carries a RUNX1/RUNX1T1 fusion, consists of cells of a specific cell stage that is distinct from the normal myeloid trajectory (Figure 2.13c). Such unique cell cluster was due to hyper CD38 level in surface protein prediction (Figure 2.13d). Such hyper-CD38 levels have been reported in AMLs with RUNX1/RUNX1T1 fusion[120-122] and recent studies have also shown that CD38 can be a potential target for adult AML[123, 124].

In this example, the imputed protein levels served as useful features for trajectory visualization. This analysis also indicates that even though cTP-net is currently trained only on

normal immune cells, it can reveal disease-specific signatures in malignant cells and the imputed protein levels are useful for characterizing tumor phenotypes.

2.3 Discussion

Taken together, our results demonstrate that cTP-net can leverage existing CITE-seq and REAP-seq datasets to predict surface protein relative abundances for new scRNA-seq data sets, and that the predictions generalize to cell types that are absent from, but related to those in the training data. cTP-net was benchmarked on PBMC and CBMC immune cells, showing good generalization across tissues and technical protocols. On Human Atlas Data, we show that the imputed surface protein levels allow easy assignment of cells to known cell types, as well as the revealing of intra-cell type gradients. We then demonstrate that, even though cTP-net used only immune cells from healthy individuals for training, it is able to impute immunophenotypes for malignant cells from acute myeloid leukemia, and that these immunophenotypes allow placement of the cells along the myeloid differentiation trajectory. Furthermore, we show that cTP-net is able to impute protein signatures in the malignant cells that are disease specific and that are not easily detectable from the transcriptomic counts.

SAVER-X serves an important role in the training procedure of cTP-net. As shown in Table 2.4, without SAVER-X denoising, the cTP-net prediction performance retracts by 0.02 in correlation, more significant than any other parameter tweaks. This discrepancy in performance is due to: (1) SAVER-X makes use of the noise model to obtain estimates of the true RNA counts. This helps cTP-net learn the underlining relationship between true RNA counts and protein level, rather than the noisy raw counts and protein levels, which varies more across data sets and thus does not generalize well. (2) By denoising the scRNA-seq, the input for learning the RNA-protein relationship is less sparse. Manifold learning on a more continuous input space usually works better[125, 126]. (3) Comparing to other autoencoder based denoising method, SAVER-X

performs Bayesian shrinkage on top of autoencoder framework to prevent over-imputation (over-smoothing) [100, 127].

Despite these promising results, cTP-net has limitations. (1) cTP-net can only apply to count based expression input (UMI-based). CITE-seq data with TPM and RPKM expression metric is not available for testing. Thus, the prediction accuracy is unknown. (2) The generalization ability of cTP-net to unrelated cell types has limitations. Even though the final cTP-net model, trained on immune cells, has good results on immune cells from diverse settings, we have not tried to perform imputation of these immune-related markers on cells that are not of the hematopoietic lineage.

With the accumulation of publicly available CITE-seq and REAP-seq data across diverse proteins, cell types and conditions, cTP-net can be retrained to accommodate more protein targets and improve in generalization accuracy. The possibility of such cross-omic transfer learning underscores the need for more diverse multi-omic cell atlases, and demonstrate how such resources can be used to enhance future studies. The cTP-net package is available both in Python and R at https://github.com/zhouzilu/cTPnet.

2.4 Methods

2.4.1 Data sets and pre-processing

Table 2.1 summarizes the five data sets analyzed in this study: CITE-PBMC, CITE-CBMC, REAP-PBMC, HCA-CBMC and HCA-BMMC. Among these, CITE-PBMC, CITE-CBMC and REAP-PBMC have paired scRNA-seq and surface protein counts, while HCA-CBMC and HCA-BMMC have only scRNA-seq counts. For all scRNA-seq data sets, low quality gene (< 10 counts across cells) and low-quality cells (less than 200 genes detected) are removed, and the count matrix (C) for all remaining cells and genes is used as input for denoising. scRNA data denoising

was performed with SAVER-X using default parameters. Denoised counts (Λ) were further transformed with Seurat default LogNormalize function,

$$X_{ij} = log\left(\frac{\Lambda_{ij} * 10{,}000}{m_j}\right) \quad (1)$$

where Λ_{ij} is the denoised molecule count of gene i in cell j, and m_j is the sum of all molecule counts of cell j. The normalized denoised count matrix X is the training input for the subsequent multiple branch neural network. For the surface protein counts, we adopted the relative abundance transformation from Stoeckius et al.[89]. For each cell c,

$$y_c = \left[\ln\left(\frac{p_{1c}}{g(\mathbf{p}_c)}\right), \ln\left(\frac{p_{2c}}{g(\mathbf{p}_c)}\right) \ldots \ln\left(\frac{p_{dc}}{g(\mathbf{p}_c)}\right)\right] \quad (2)$$

where $\mathbf{p}_c$ is vector of antibody-derived tags (ADT) counts, and $g(\mathbf{p}_c)$ is the geometric mean of $\mathbf{p}_c$. The network trained using this transformed relative protein abundance as the response vector yields better prediction accuracy than the network trained using raw protein barcode counts.

2.4.2 cTP-net neural network structure and training parameters

Figure 2.4 shows the structure of cTP-net. Here, we have a normalized expression matrix $\mathbf{X}$ of N cells and D genes, and a normalized protein abundance matrix $\mathbf{Y}$ of the same N cells and d surface proteins. Let's denote cTP-net as a function F that maps from $\mathbb{R}^D$ to $\mathbb{R}^d$. Starting from the input layer, with dimension equals to number of genes D, the first internal layer has dimension 1000, followed by a second internal layer with dimension 128. These two layers are designed to learn and encode features that are shared across proteins, such as features that are informative for cell type, cell state and common processes such as cell cycle. The remaining layers are protein specific, with 64 nodes for each protein that feed into a one node output layer giving the imputed value. All layers except the last layer are fully connected (FC) with rectified linear unit

(ReLU) activation function [128], while the last layer is a fully connected layer with identity activation function for output. The objective function here is,

$$\underset{F}{\operatorname{argmin}}|\mathbf{Y} - F(\mathbf{X})|_1 \quad (3)$$

where the loss is L1 norm. The objective function was optimized stochastically with Adam [129] with learning rate set to 10e-5 for 139 epochs (cross-validation). Other variations of cTP-net, which we found to have inferior performance, are illustrated in more details in Table 2.4. The first column indicates the differences to the finalized models, while the second column shows the correlation of the predicted protein abundance to the true protein abundance in the holdout setting on CITE-seq CBMC data set. As shown by Table 2.4, missing any component of the final model will result in inferior performance.

2.4.3 Benchmarking procedure

Figure 2.5a shows the validation set testing procedure. Given limited amount of data, we keep only 10% of the cells as the testing set, and use the other 90% of the cells for training. The optimal model was selected based on the testing error.

We perform the out-of-cell type prediction based on Figure 2.5b. This procedure mimics cross-validation, except that, instead of selecting the test set cells randomly, we partition the cells by their cell types. Iteratively, we designate all cells of a given cell type for testing and use the remaining cells for training. We then perform prediction on the hold-out cell type using the model trained on all other cell types. In the end, every cell has been tested once and has the corresponding predictions. In the benchmark against the validation set testing procedure, we limit comparisons to the same cells that were in the validation set in the holdout scheme to account for variations between subsets.

To apply the models we trained in validation set testing procedure to different cell populations and technologies, the inputs have to be in the same feature space. Even though all

data sets considered are from human cells, the list of genes differs between experiments and technologies. Genes that are in the training data but not in the testing data are filled with zeros. Because cTP-net utilizes overrepresented number of genes to predict the surface proteins level, having a small number of genes missing has little effect on the performance. After prediction, we selected only the shared proteins between two data sets for comparison.

2.4.4 cTP-net interpolation

To better interpret the relationships that the neural network is learning, we developed a permutation-based interpolation scheme that can calculate an influence score epi for each gene in the imputation of each protein (Figure 2.8). The idea is to assess how much changing the expression value of certain genes in the training data affects the training errors for a given model F. In each epoch, we interpolate all of the genes in a stochastic manner. Let's denote $\mathbf{X}$ as the expression matrix (N by G matrix, where N is the number of cells and G is number of genes), $\mathbf{Y}$ as protein abundance matrix and L as the loss function. The algorithm goes as follow (Figure 2.8):

(1) Estimate the original model error $\epsilon^{orig} = L(\mathbf{Y}, F(\mathbf{X}))$.

(2) Sampling batch of genes denote by . Generate expression matrix $\mathbf{X}^{perm}$ by permuting genes in gs in the data $\mathbf{X}$. This breaks the association between gs and protein abundance $\mathbf{Y}$, i.e. the cell order within gs does not coordinate with protein abundance $\mathbf{Y}$.

(3) Estimate error $\epsilon^{perm} = L(\mathbf{Y}, F(\mathbf{X}^{perm}))$ based on the predictions of the permuted data.

(4) Calculate permutation feature importance $\Delta_{gs} = |\epsilon^{orig} - \epsilon^{perm}|$ of gene set gs to this model F.

We set batch size as 100 with 500 epochs. Furthermore, by picking different cells to interpolate, we could identify gene influence score in different cell types. For example, if matrix $\mathbf{X}$ belongs to a given cell type, the cell type specific genes are consistent across cells of the given cell type, and thus, the permutation will not influence these genes. Genes that influence the

surface protein abundance within the cell type, such as cell cycle genes and protein synthesis genes, tend to be rewarded with high influence scores in such a cell-type specific interpolation analysis.

For the top 100 highest influence scored genes from the following scenarios in CITE-PBMC: (1) CD45RA in CD14-CD16+ monocytes, (2) CD11c in CD14-CD16+ monocytes, (3) CD45RA in CD8 T cells, (4) CD45RA in CD4 T cells, (5) CD11c in CD14+CD16+ monocytes, (6) CD45RA in dendritic cells, and (7) CD11c in dendritic cells, we employed a Gene Ontology analysis [57] which identify top 10 pathways based on GO gene sets with FDR q-value < 0.05 as significant (Table 2.6).

2.4.5 Seurat anchor-transfer analysis

We compared cTP-net with an anchor-based transfer learning method developed in Seurat v3 [104]. For Seurat v3, RNA count data are normalized by LogNormalization method, while surface protein counts are normalized by centered log-ratio (CLR) method. In validation test setting, we used the same cells for training and testing as in cTP-net so as to be directly comparable to cTP-net. For out-of-cell type prediction, default parameters did not work for several cell types in anchor-transfer step, because, for those cell types, there are few anchors shared between the training and testing sets. To overcome this, we reduced the number of anchors iteratively until the function ran successfully.

2.4.6 HCA data analysis

HCA RNA-seq data sets are pre-processed as discussed above, resulting in log-normalized denoised values. We applied default pipeline of Seurat and generated UMAP plot for both data sets (Figure 2.15). Cells are clearly clustered by individuals, indicating strong batch effects. As a result, the following analysis was performed on cells of each individual. Major cell types were determined by known markers.

From the log-normalized denoised expression value, we predict the surface protein abundance with cTP-net model trained jointly on CITE-seq PBMC, CBMC and BMMC data sets. We embedded 24 surface protein abundance across 16 individuals on t-SNE plot, showing consistent results with cell type information (Figure 2.10, 2.11).

2.5 Data availability

Public datasets for training and evaluating cTP-net can be found at National Center for Biotechnology Information Gene Expression Omnibus (GEO) under accession number GSE100866, GSE100501 and GSE128639 respectively.

2.6 Code availability

cTP-net package are publicly available as both an open-source R package at https://github.com/zhouzilu/cTPnet with license GPL-3.0 and an open-source python package at https://github.com/zhouzilu/ctpnetpy with license GPL-3.0.

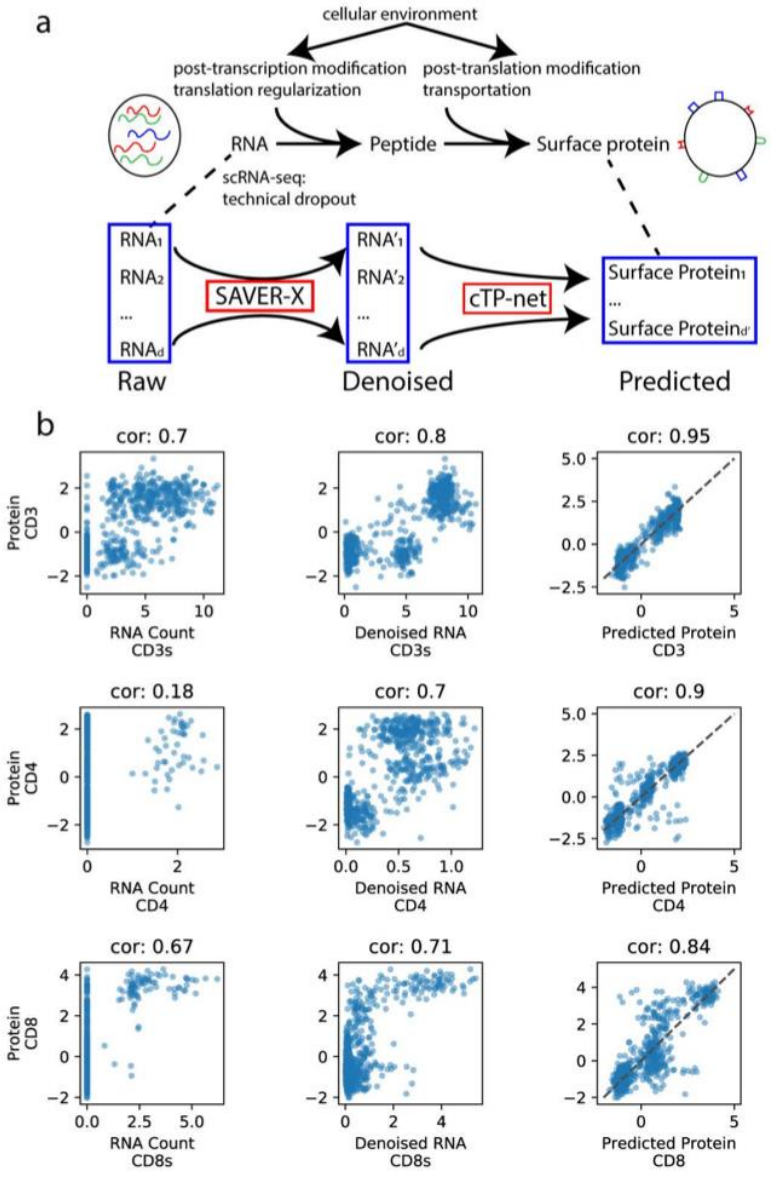
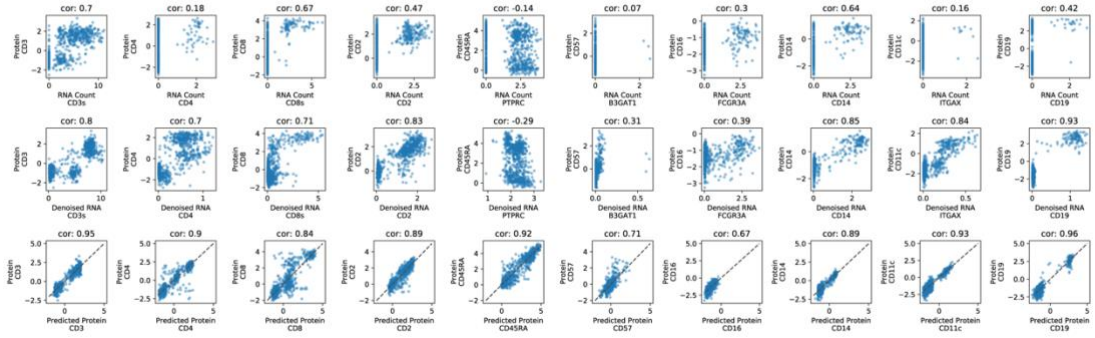
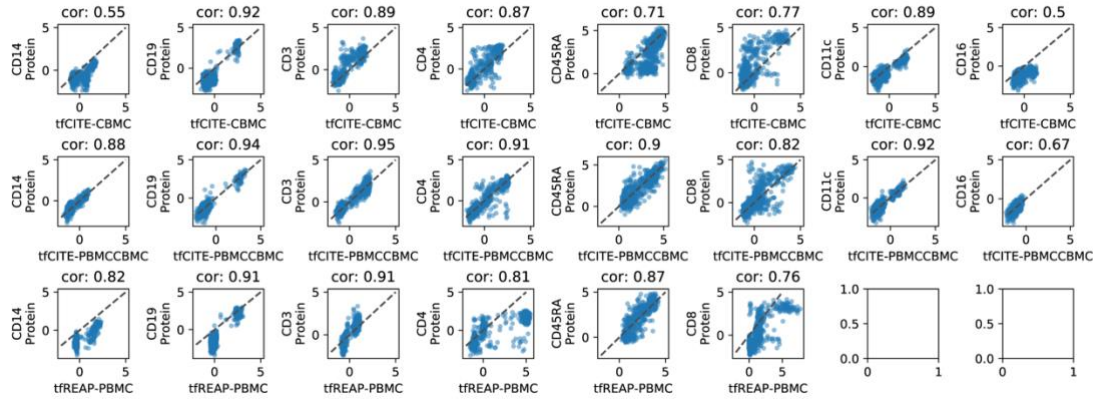


Figure 2.1 ***cTP-net analysis pipeline and imputation of example proteins.*** **(a)** *Overview of cTP-net analysis pipeline, which learns a mapping from the denoised scRNA-seq data to the relative abundance of surface proteins, capturing multi-gene features that reflect the cellular environment and related processes.* ***(b)*** *For three example proteins (CD3, CD4 and CD8), cross-cell scatter and correlation (cor) of CITE-seq measured abundances vs. (1) raw RNA count ("CD3s" and "CD8s" are sum of all genes that compose protein CD3 and CD8, see Table 2.5), (2) SAVER-X denoised RNA level, and (3) cTP-net predicted protein abundance.*

a


cor: 0.7
cor: 0.18
cor: 0.67
cor: 0.47
cor: -0.14
cor: 0.07
cor: 0.3
cor: 0.64
cor: 0.16
cor: 0.42
cor: 0.8
cor: 0.7
cor: 0.71
cor: 0.83
cor: -0.29
cor: 0.31
cor: 0.39
cor: 0.85
cor: 0.84
cor: 0.93
cor: 0.95
cor: 0.9
cor: 0.84
cor: 0.89
cor: 0.92
cor: 0.71
cor: 0.67
cor: 0.89
cor: 0.93
cor: 0.96

b


cor: 0.55
cor: 0.92
cor: 0.89
cor: 0.87
cor: 0.71
cor: 0.77
cor: 0.89
cor: 0.5
cor: 0.88
cor: 0.94
cor: 0.95
cor: 0.91
cor: 0.9
cor: 0.82
cor: 0.92
cor: 0.67
cor: 0.82
cor: 0.91
cor: 0.91
cor: 0.81
cor: 0.87
cor: 0.76

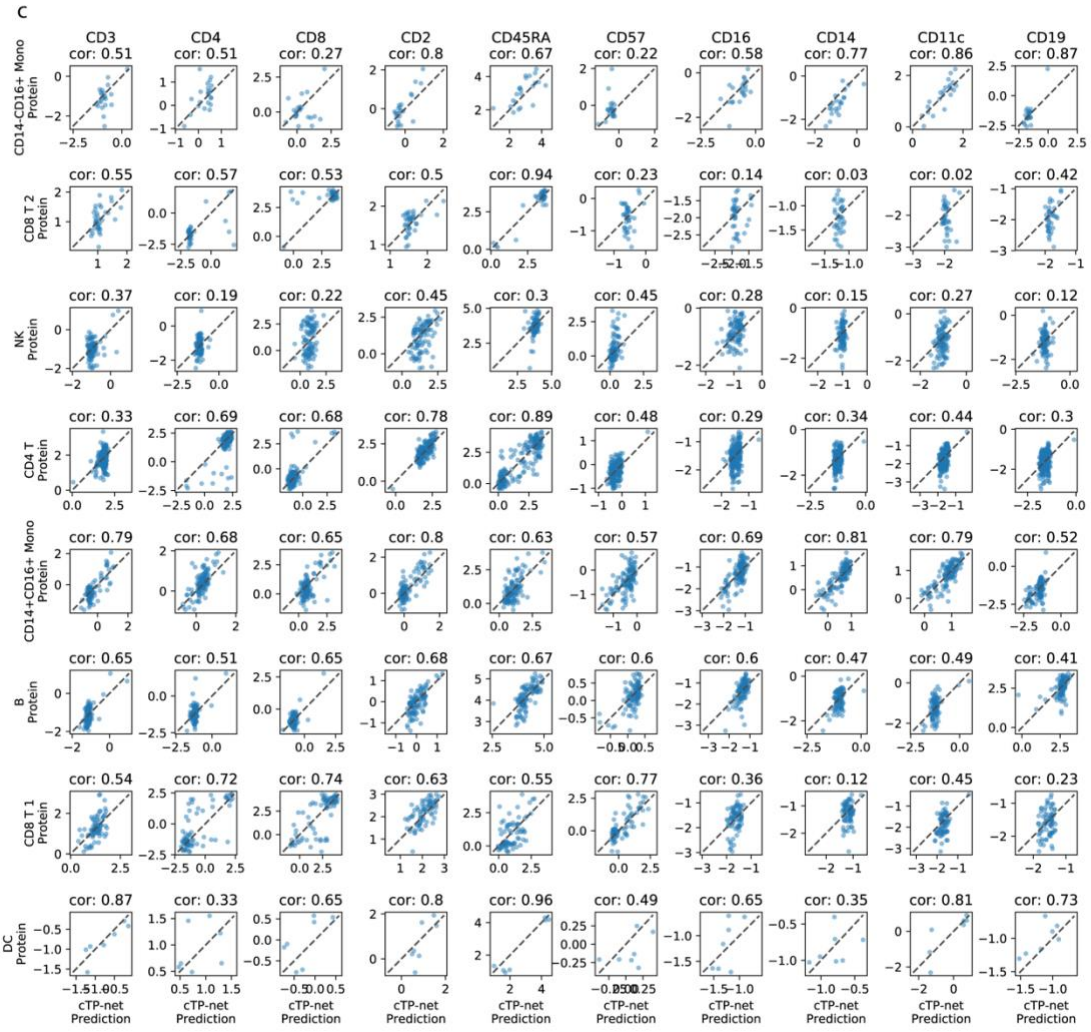
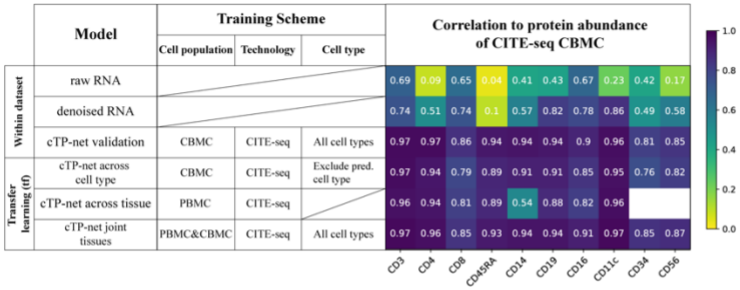
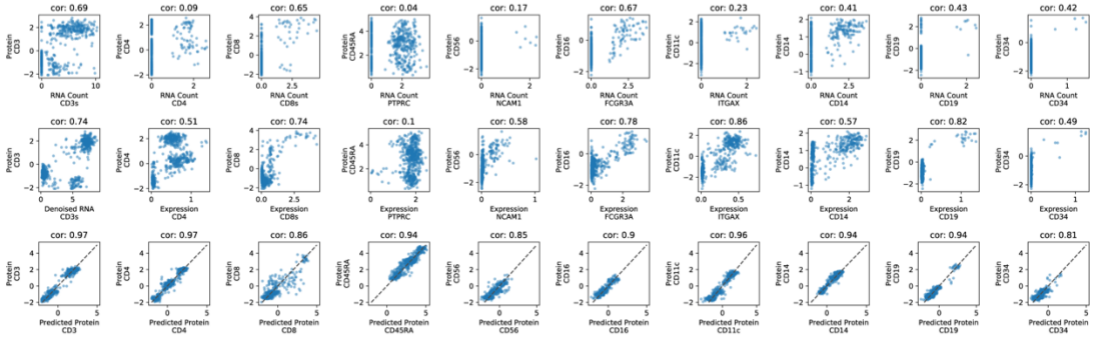


Figure 2.2 ***Benchmark evaluation of cTP-net on CITE-PBMC data set. (a)*** *Benchmark correlation of true protein level vs. (1) Raw RNA count, (2) SAVER-X denoised RNA level, and (3) cTP-net predicted protein abundance in holdout method.* ***(b)*** *Benchmark correlation of truth protein level vs. (1) transfer learning from CITE-CBMC,* ***(2)*** *transfer learning from CITE-PBMCCBMC, and (3) transfer learning from REAP-PBMC. (c) Benchmark correlation of true protein level vs. cTP-net prediction in holdout method for each cell type.*

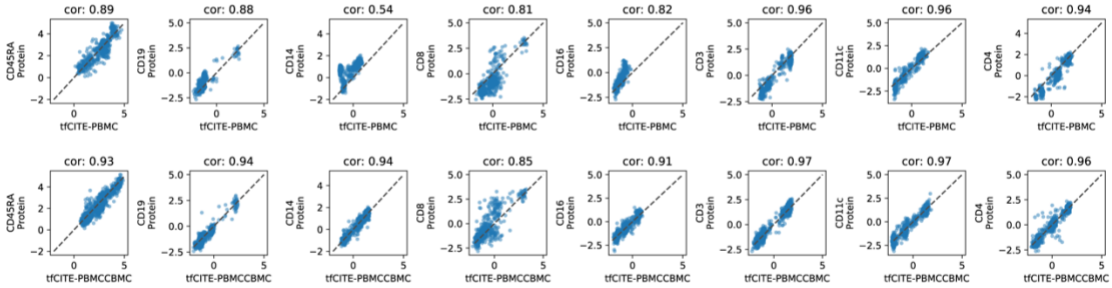
a


Model
Training Scheme
Cell population
Technology
Cell type
Correlation to protein abundance of CITE-seq CBMC
Within dataset
Transfer learning (tf)
raw RNA
0.69 0.09 0.65 0.04 0.41 0.43 0.67 0.23 0.42 0.17
denoised RNA
0.74 0.51 0.74 0.1 0.57 0.82 0.78 0.86 0.49 0.58
cTP-net validation
CBMC CITE-seq All cell types
0.97 0.97 0.86 0.94 0.94 0.94 0.9 0.96 0.81 0.85
cTP-net across cell type
CBMC CITE-seq Exclude pred. cell type
0.97 0.94 0.79 0.89 0.91 0.91 0.85 0.95 0.76 0.82
cTP-net across tissue
PBMC CITE-seq
0.96 0.94 0.81 0.89 0.54 0.88 0.82 0.96
cTP-net joint tissues
PBMC&CBMC CITE-seq All cell types
0.97 0.96 0.85 0.93 0.94 0.94 0.91 0.97 0.85 0.87
CD3 CD4 CD8 CD45RA CD14 CD19 CD16 CD11c CD34 CD56
1.0 0.8 0.6 0.4 0.2 0.0

b


cor: 0.69 cor: 0.09 cor: 0.65 cor: 0.04 cor: 0.17 cor: 0.67 cor: 0.23 cor: 0.41 cor: 0.43 cor: 0.42
cor: 0.74 cor: 0.51 cor: 0.74 cor: 0.1 cor: 0.58 cor: 0.78 cor: 0.86 cor: 0.57 cor: 0.82 cor: 0.49
cor: 0.97 cor: 0.97 cor: 0.86 cor: 0.94 cor: 0.85 cor: 0.9 cor: 0.96 cor: 0.94 cor: 0.94 cor: 0.81
CD3 CD4 CD8 CD45RA CD56 CD16 CD11c CD14 CD19 CD34
RNA Count
Denoised RNA
Expression
Predicted Protein

c


cor: 0.89 cor: 0.88 cor: 0.54 cor: 0.81 cor: 0.82 cor: 0.96 cor: 0.96 cor: 0.94
cor: 0.93 cor: 0.94 cor: 0.94 cor: 0.85 cor: 0.91 cor: 0.97 cor: 0.97 cor: 0.96
CD45RA Protein
CD19 Protein
CD14 Protein
CD8 Protein
CD16 Protein
CD3 Protein
CD11c Protein
CD4 Protein
tfCITE-PBMC
tfCITE-PBMCCBMC

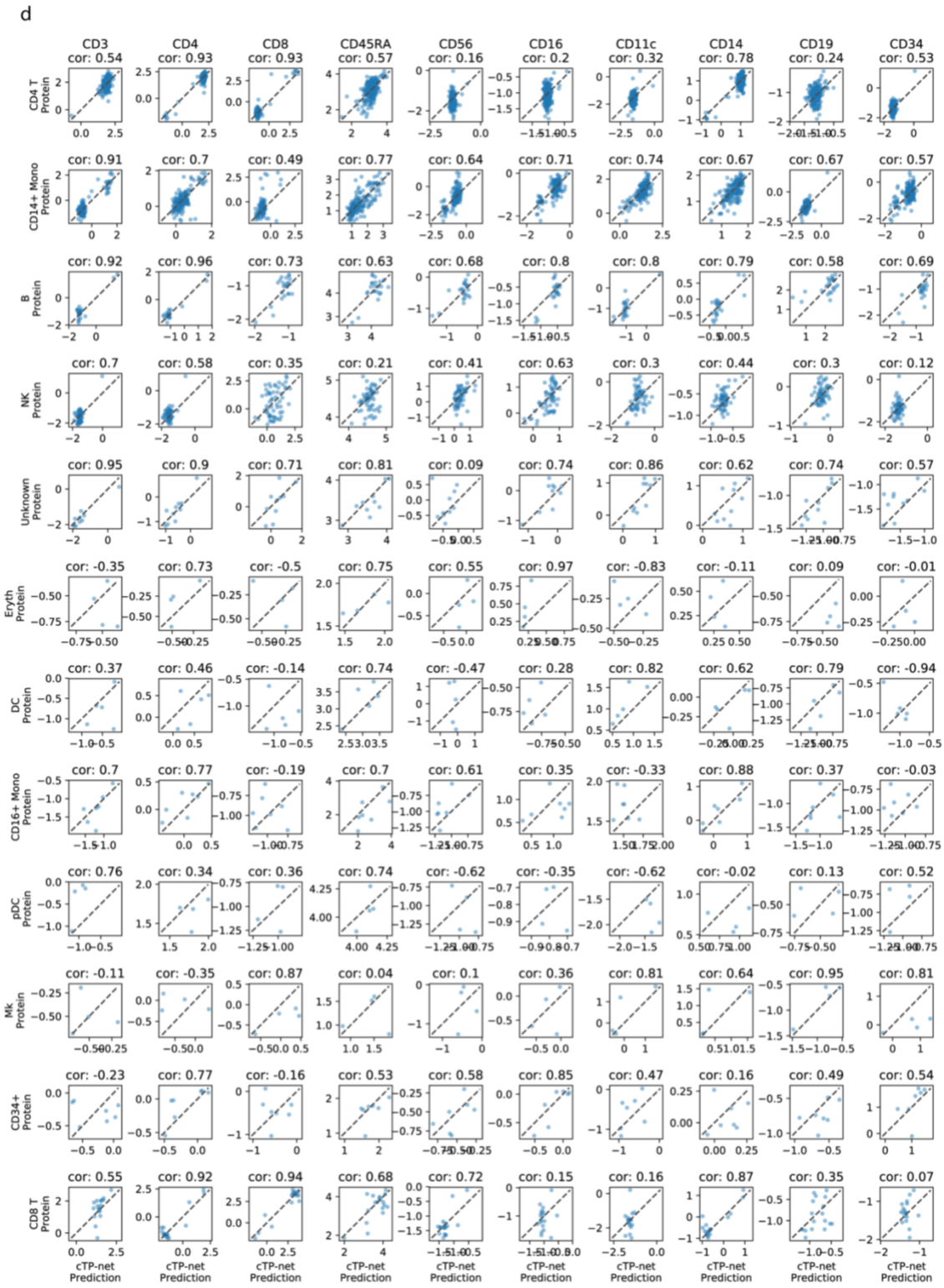

d
CD3
CD4
CD8
CD45RA
CD56
CD16
CD11c
CD14
CD19
CD34
CD4 T Protein
CD14+ Mono Protein
B Protein
NK Protein
Unknown Protein
Eryth Protein
DC Protein
CD16+ Mono Protein
pDC Protein
Mk Protein
CD34+ Protein
CD8 T Protein
cTP-net Prediction

Figure 2.3 ***Benchmark evaluation of cTP-net on CITE-CBMC data set. (a)*** *Benchmark evaluation heatmap of cTP-net and comparison with Seurat v3. The table on the left captures the detailed training scheme and model name of each test.* ***(b)*** *Benchmark correlation of true protein level vs. (1) Raw RNA count, (2) SAVER-X denoised RNA level, and (3) cTP-net predicted protein abundance in holdout method.* ***(c)*** *Benchmark correlation of truth protein level vs. (1) transfer learning from CITE-PBMC, and (2) transfer learning from CITE-PBMCCBMC.* ***(d)*** *Benchmark correlation of true protein level vs. cTP-net prediction in holdout method for each cell type.*

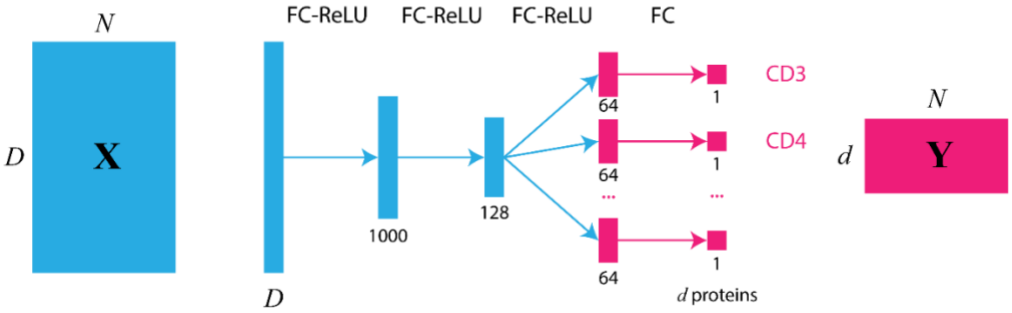


Figure 2.4 ***Neural network architecture of the cTP-net.***

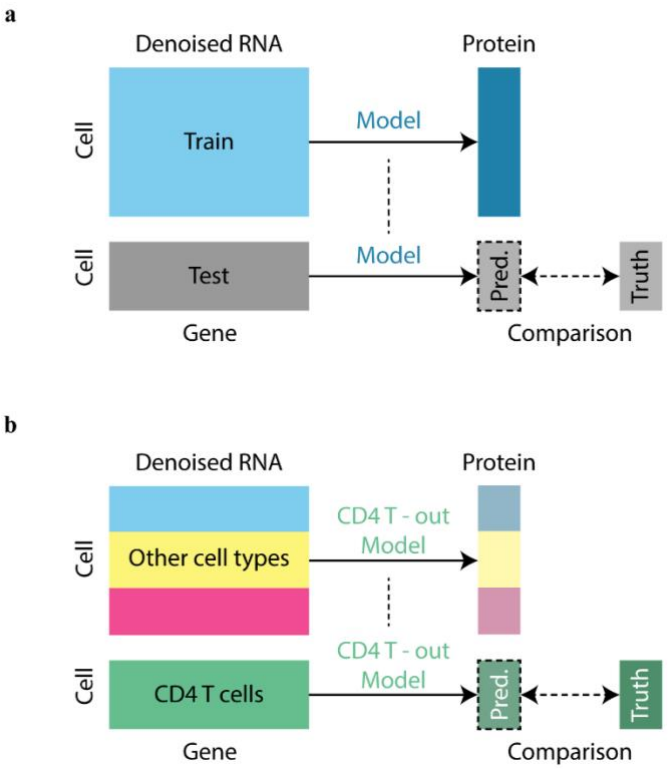


Figure 2.5 ***Benchmark procedure. (a)*** *Holdout method validation scheme.* ***(b)*** *Out-of-cell-type benchmark scheme.*

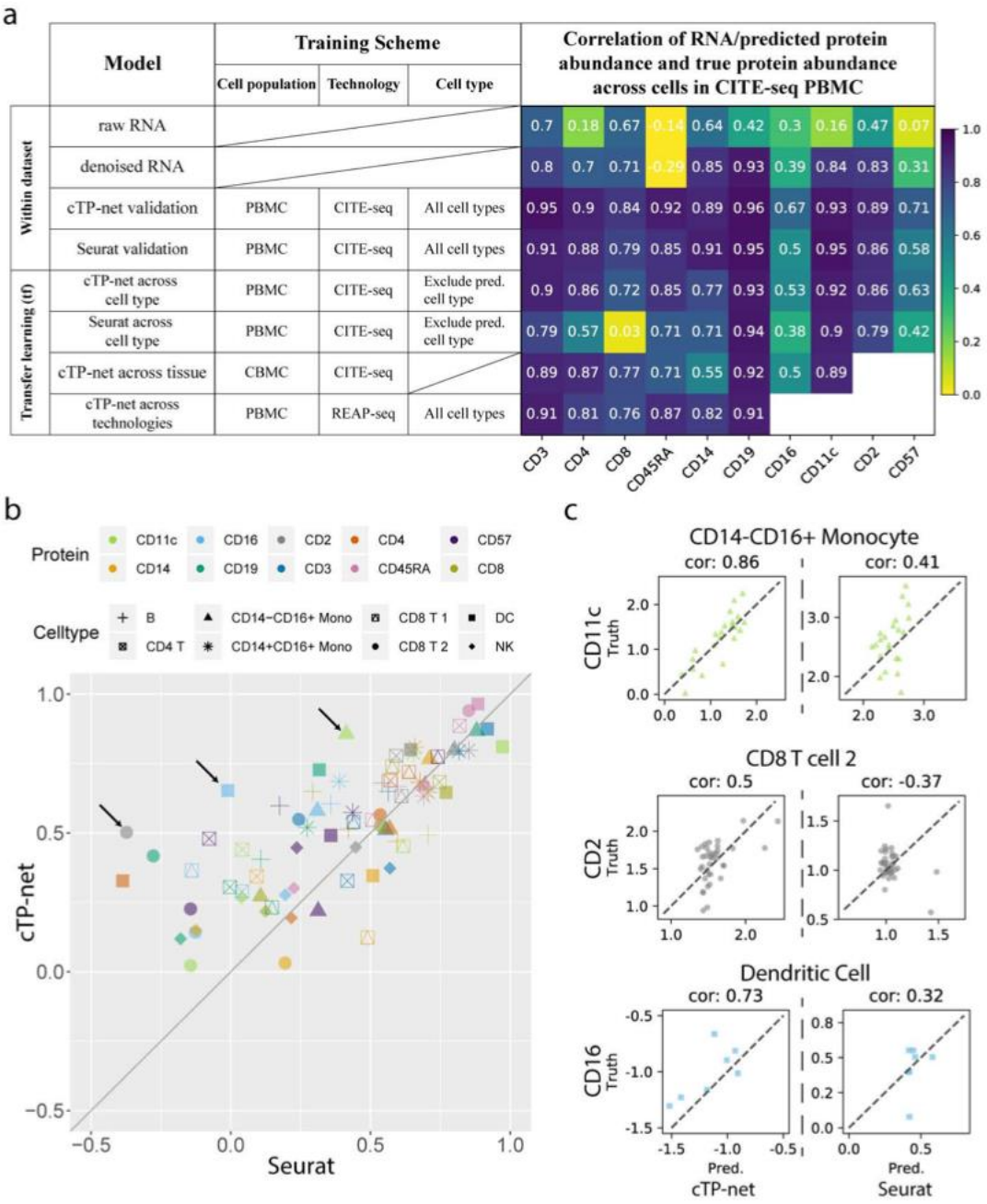
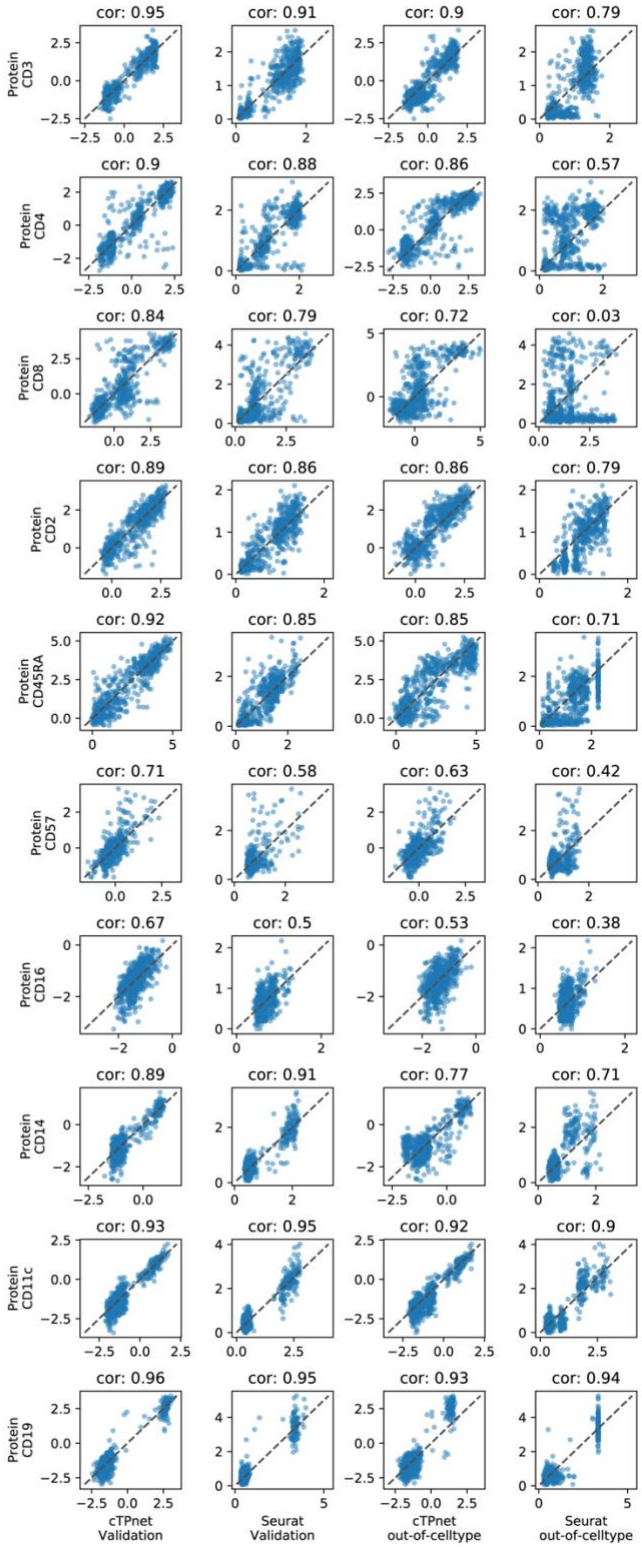


Figure 2.6 ***Benchmark evaluation on CITE-seq PBMC data. (a)*** *Benchmark evaluation of cTP-net on CITE-seq PBMC data, with comparisons to Seurat v3, in validation, across cell type, across tissue and across technology scenarios. The table on the left shows the training scheme of each test, the heatmap shows correlations with actual measured protein abundances.* ***(b)***

Within cell type correlations between imputed and measured protein abundance on the CITE-seq PBMC data, Seurat v3 versus cTP-net. Each point (color and shape pair) indicates a cell type and surface protein pair, where the x-axis is correlation between actual measured abundance and Seurat imputation and y-axis is the correlation between actual measured abundance and cTP-net imputation. ***(c)*** *Scatter of imputed versus measured abundance for the three (surface protein, cell type) pairs marked by arrows in (b): CD11c in CD14-CD16+ monocytes, CD2 in CD8 T cells, and CD19 in dendritic cells.*

a


cor: 0.95
cor: 0.91
cor: 0.9
cor: 0.79
Protein CD3
cor: 0.9
cor: 0.88
cor: 0.86
cor: 0.57
Protein CD4
cor: 0.84
cor: 0.79
cor: 0.72
cor: 0.03
Protein CD8
cor: 0.89
cor: 0.86
cor: 0.86
cor: 0.79
Protein CD2
cor: 0.92
cor: 0.85
cor: 0.85
cor: 0.71
Protein CD45RA
cor: 0.71
cor: 0.58
cor: 0.63
cor: 0.42
Protein CD57
cor: 0.67
cor: 0.5
cor: 0.53
cor: 0.38
Protein CD16
cor: 0.89
cor: 0.91
cor: 0.77
cor: 0.71
Protein CD14
cor: 0.93
cor: 0.95
cor: 0.92
cor: 0.9
Protein CD11c
cor: 0.96
cor: 0.95
cor: 0.93
cor: 0.94
Protein CD19
cTPnet Validation
Seurat Validation
cTPnet out-of-celltype
Seurat out-of-celltype

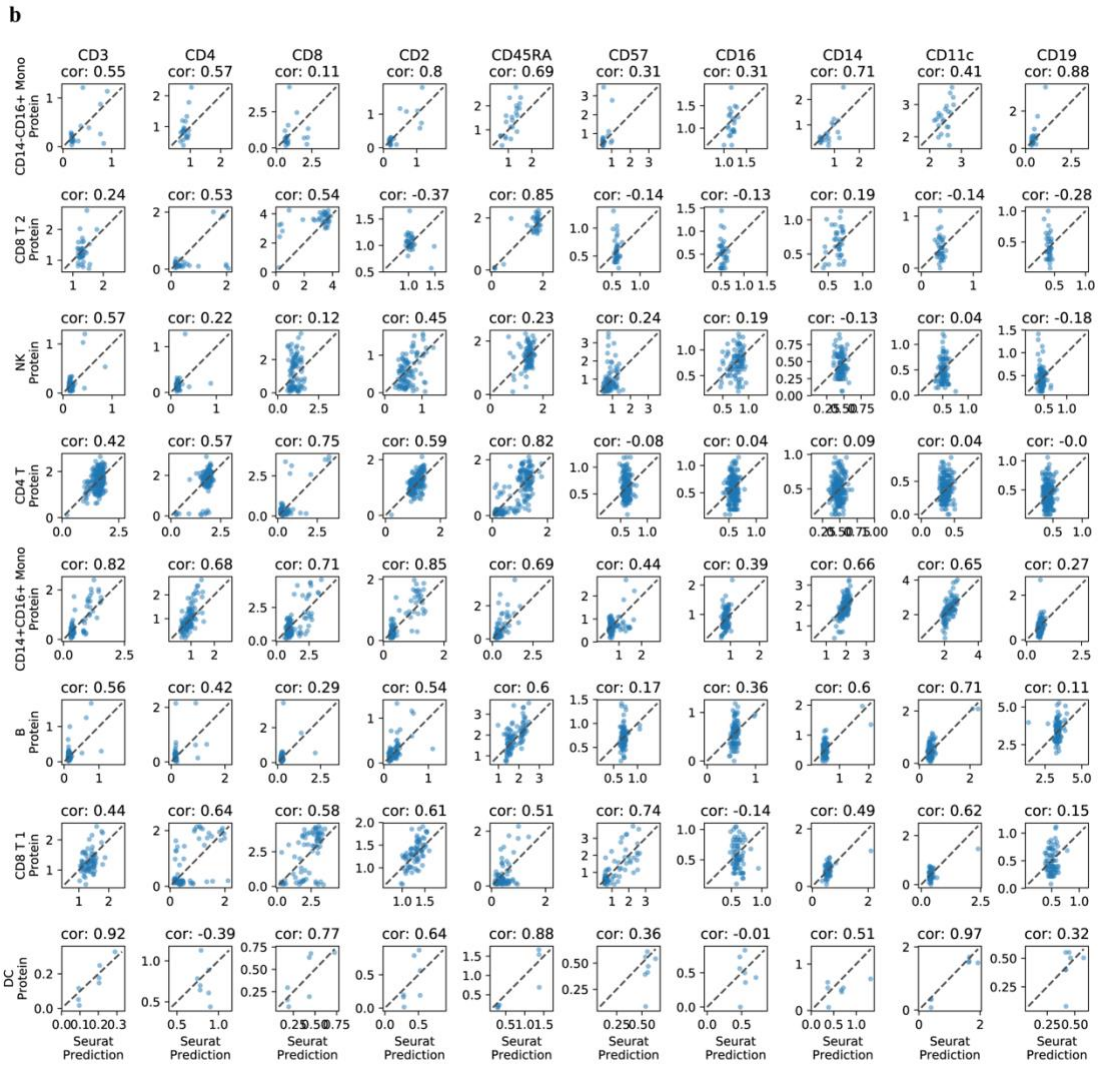


Figure 2.7 ***Benchmark evaluation of Seurat v3 on CITE-PBMC data set. (a)*** *Benchmark correlation of true protein level vs. (1) cTP-net predicted protein abundance in holdout method, (2) Seurat v3 predicted protein abundance in holdout method, (3) out-of-cell-type cTP-net predicted protein abundance, and (4) out-of-cell-type Seurat v3 predicted protein abundance.* ***(b)*** *Benchmark correlation of truth protein level vs. (1) transfer learning from CITE-PBMC, and (2) transfer learning from CITE-PBMCCBMC.* ***(c)*** *Benchmark correlation of true protein level vs. cTP-net prediction in holdout method for each cell type.*

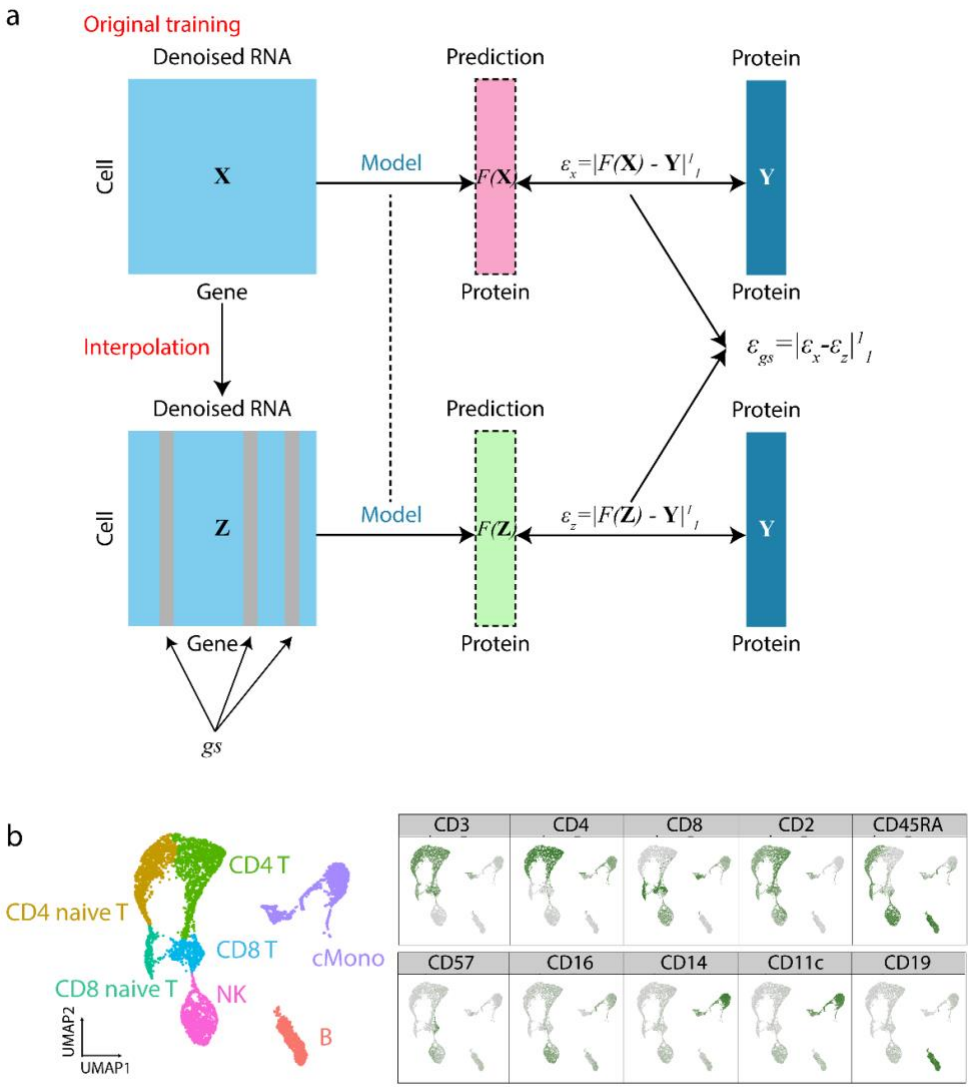


Figure 2.8 ***Interpolation analysis. (a)*** *Interpolation procedure in identify permutation based importance score for each gene in each protein prediction.* ***(b)*** *Dimension reduction analysis on the bottleneck layer on cTP-net trained on PBMCs from CITE-seq.*

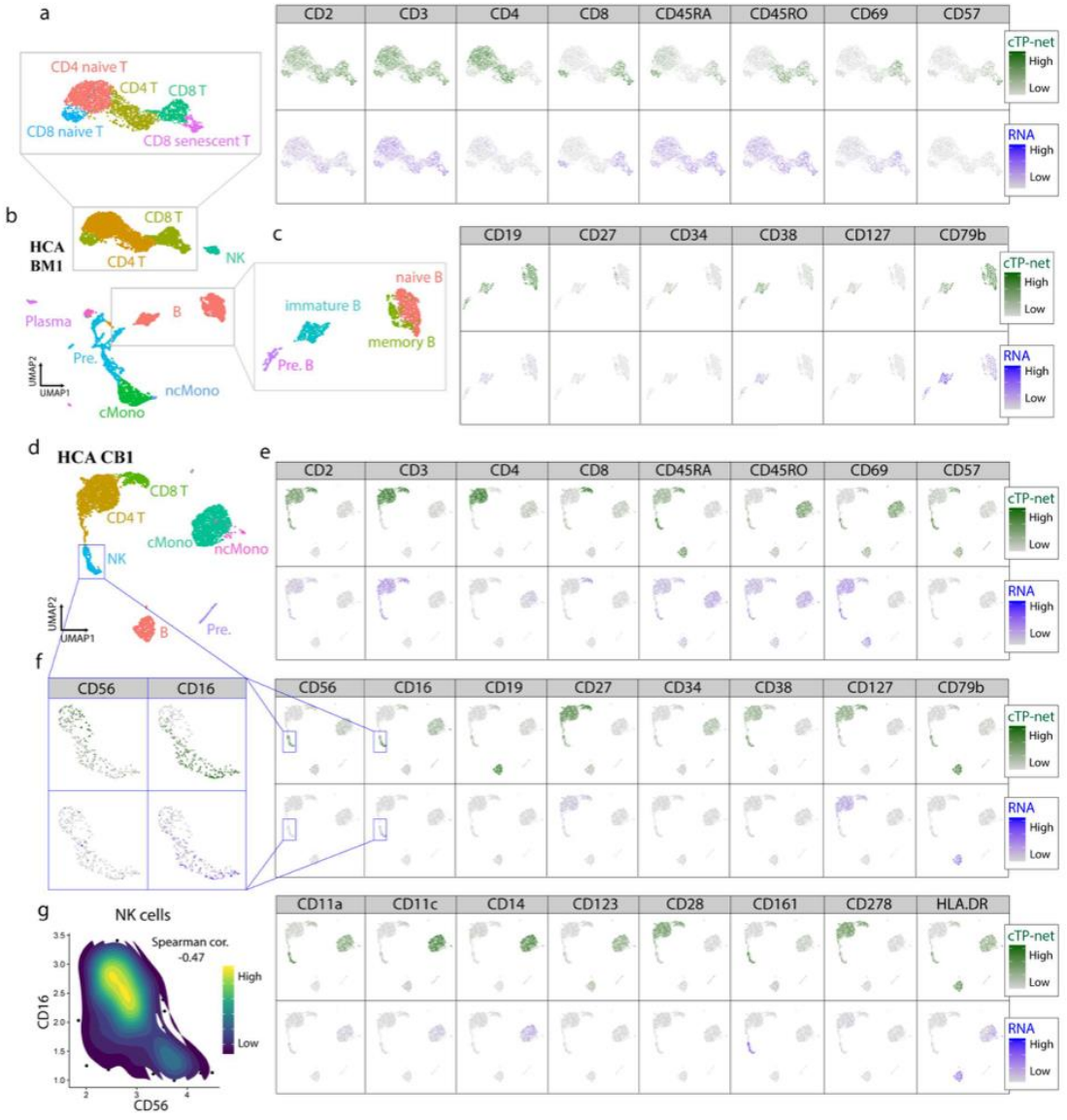
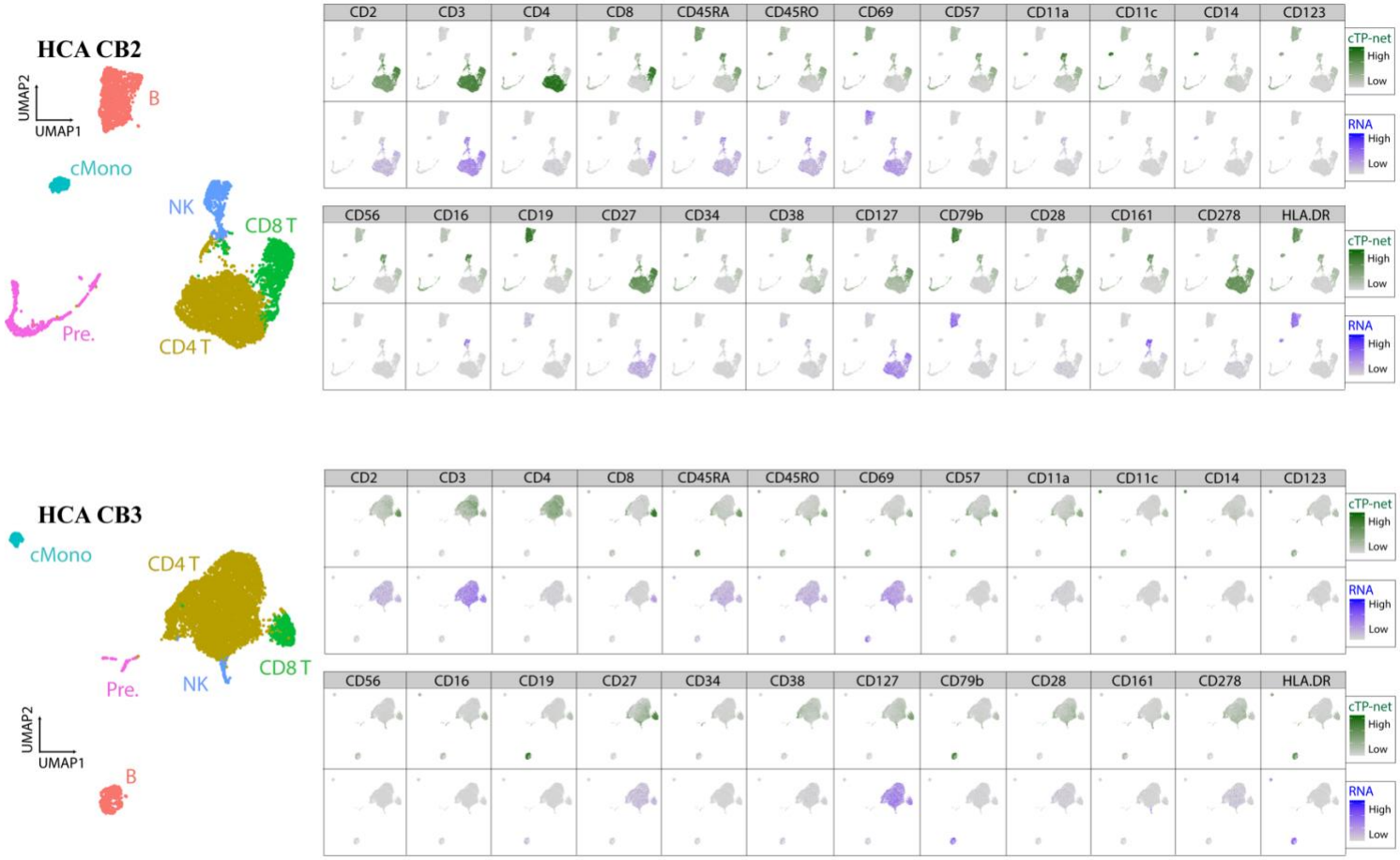
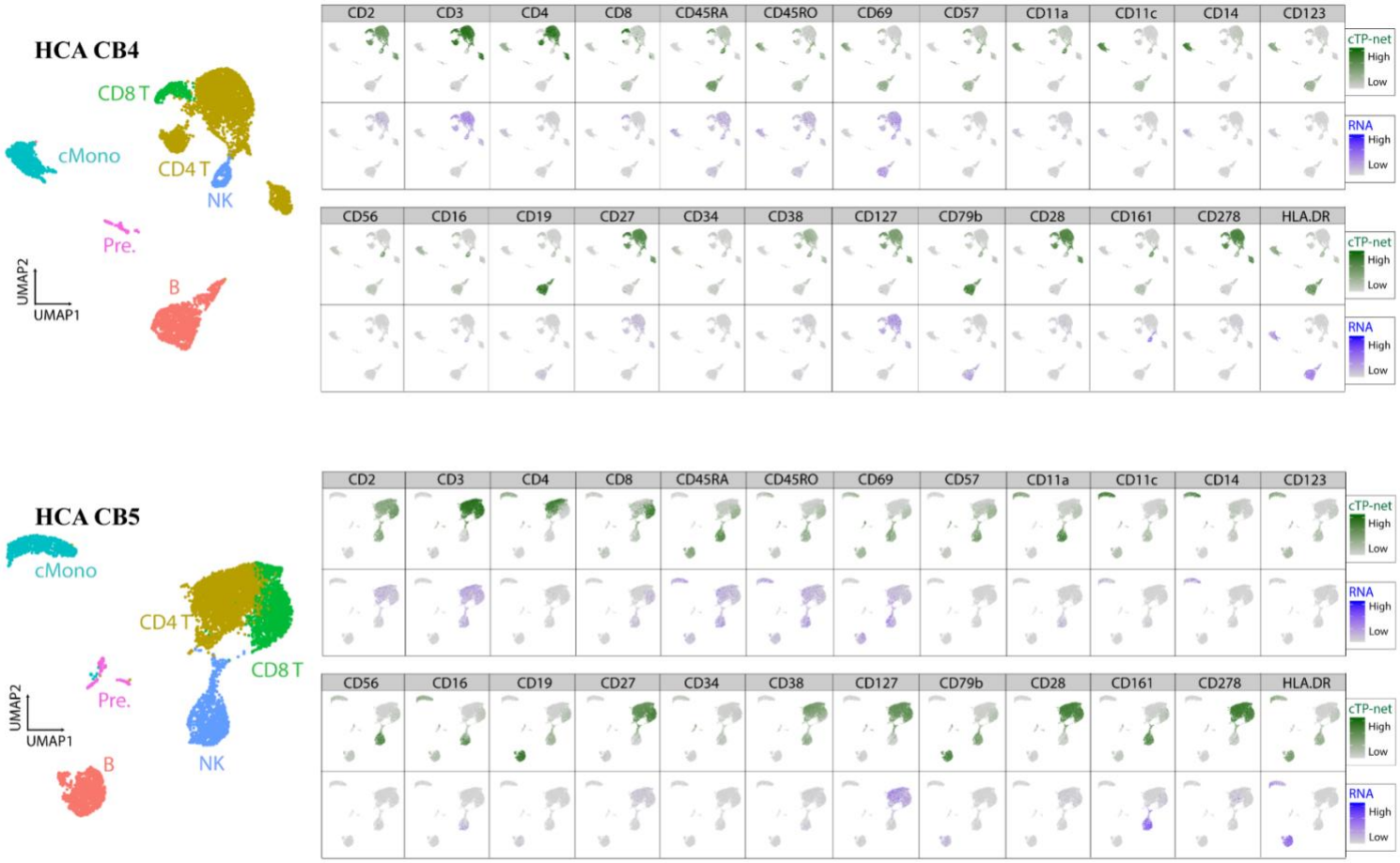
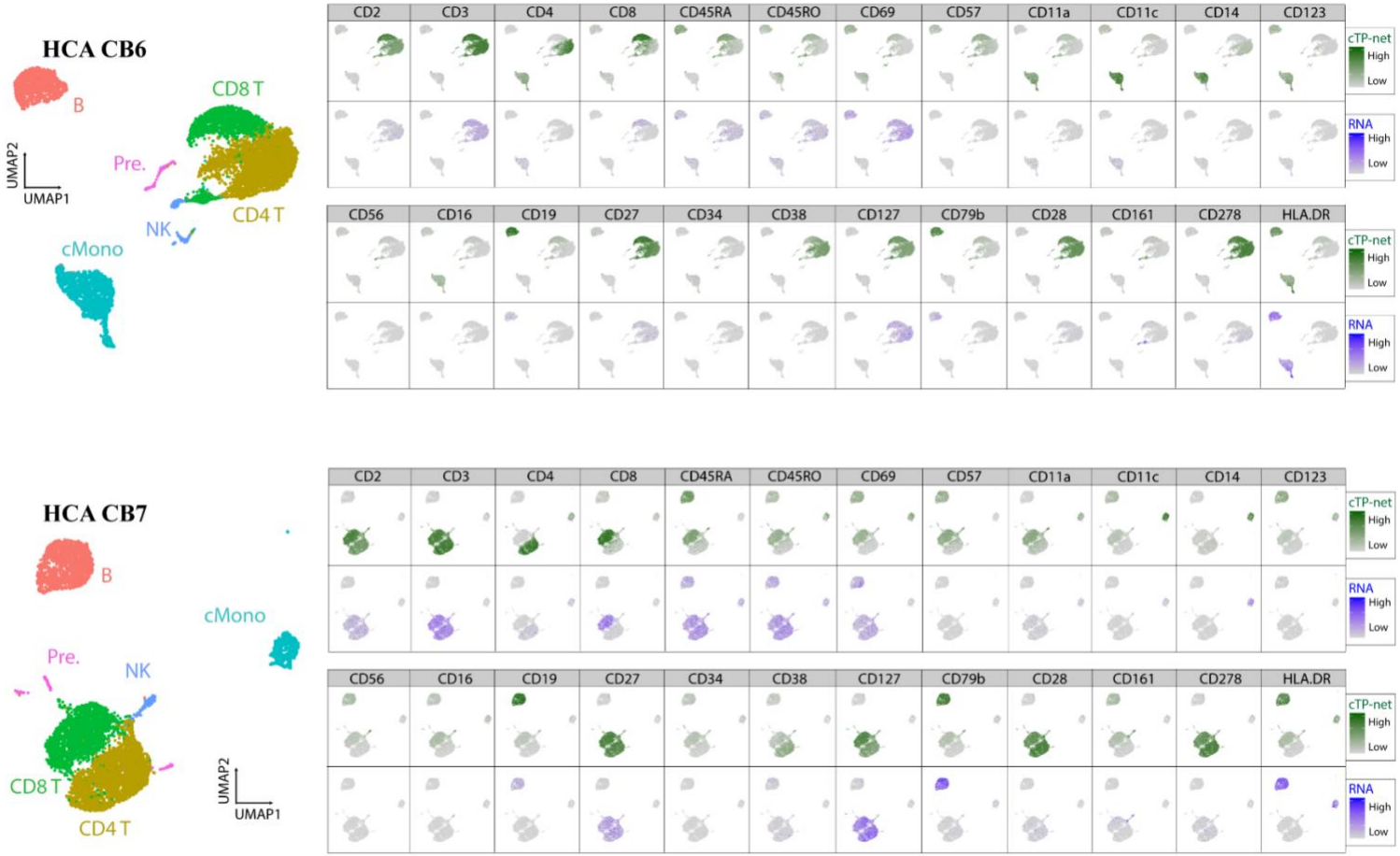


Figure 2.9 ***Imputation results analysis on Human Cell Atlas data sets. (a)*** *Left panel: UMAP visualization of MantonBM1 BMMCs T cell subpopulation based on RNA expression, colored by cell type. CD4 T: mature CD4+ T cells; mature CD8 T: CD8+ T cells; naïve CD4 T: naïve CD4+ T cells; naïve CD8 T: naïve CD8+ T cells; CD8 senescent T: CD8+ senescent T cells. Right panel: Related imputed protein abundance and RNA expression of its corresponding gene.* ***(b)*** *UMAP visualization of MantonBM1 BMMCs based on RNA expression, colored by cell type. B: B cells; CD4 T: CD4+ T cells; CD8 T: CD8+ T cells; cMono: classical monocyte; ncMono: non-classical*

monocyte; NK: natural killer cells; Pre.: precursors; Plasma: plasma cells. ***(c)*** *Left panel: UMAP visualization of MantonBM1 BMMCs B cell subpopulation based on RNA expression, colored by cell type. Pre.B: B cell precursors; immature B: immature B cells; memory B: memory B cells; naïve B: naïve B cells. Right panel: Related imputed protein abundance and RNA expression of its corresponding gene.* ***(d)*** *UMAP visualization of MantonCB2 CBMCs based on RNA expression, colored by cell type.* ***(e)*** *cTP-net imputed protein abundance and RNA read count of its corresponding gene for 24 surface proteins.* ***(f)*** *UMAP visualization of MantonCB2 CBMCs NK cell subpopulation colored by CD56 and CD16 imputed protein abundance and RNA read count. Reverse gradient is observed in cTP-net prediction but not in the read count for its corresponding RNA.* ***(g)*** *Contour plot of cells based on imputed CD56 and CD16 abundance in NK cell populations. Strong negative correlation (Spearman correlation = -0.47) with two subpopulation observed.*


HCA CB2
UMAP2
UMAP1
B
cMono
NK
CD8 T
Pre.
CD4 T
CD2
CD3
CD4
CD8
CD45RA
CD45RO
CD69
CD57
CD11a
CD11c
CD14
CD123
CD56
CD16
CD19
CD27
CD34
CD38
CD127
CD79b
CD28
CD161
CD278
HLA.DR
cTP-net
High
Low
RNA
High
Low
HCA CB3
cMono
CD4 T
Pre.
NK
CD8 T
UMAP2
UMAP1
B


HCA CB4
CD8 T
cMono
CD4 T
NK
Pre.
B
UMAP2
UMAP1
CD2
CD3
CD4
CD8
CD45RA
CD45RO
CD69
CD57
CD11a
CD11c
CD14
CD123
CD56
CD16
CD19
CD27
CD34
CD38
CD127
CD79b
CD28
CD161
CD278
HLA.DR
cTP-net
High
Low
RNA
High
Low
HCA CB5
cMono
CD4 T
CD8 T
Pre.
NK
B
UMAP2
UMAP1


HCA CB6
B
CD8 T
Pre.
CD4 T
NK
cMono
UMAP2
UMAP1
CD2
CD3
CD4
CD8
CD45RA
CD45RO
CD69
CD57
CD11a
CD11c
CD14
CD123
CD56
CD16
CD19
CD27
CD34
CD38
CD127
CD79b
CD28
CD161
CD278
HLA.DR
cTP-net
High
Low
RNA
High
Low
HCA CB7
B
cMono
Pre.
NK
CD8 T
CD4 T
UMAP2
UMAP1

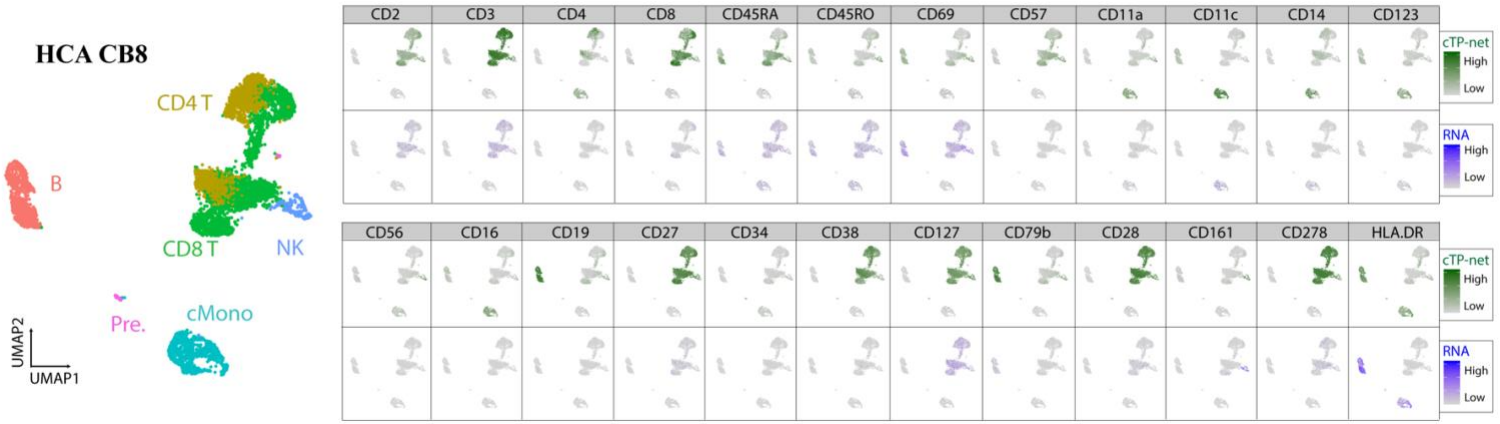
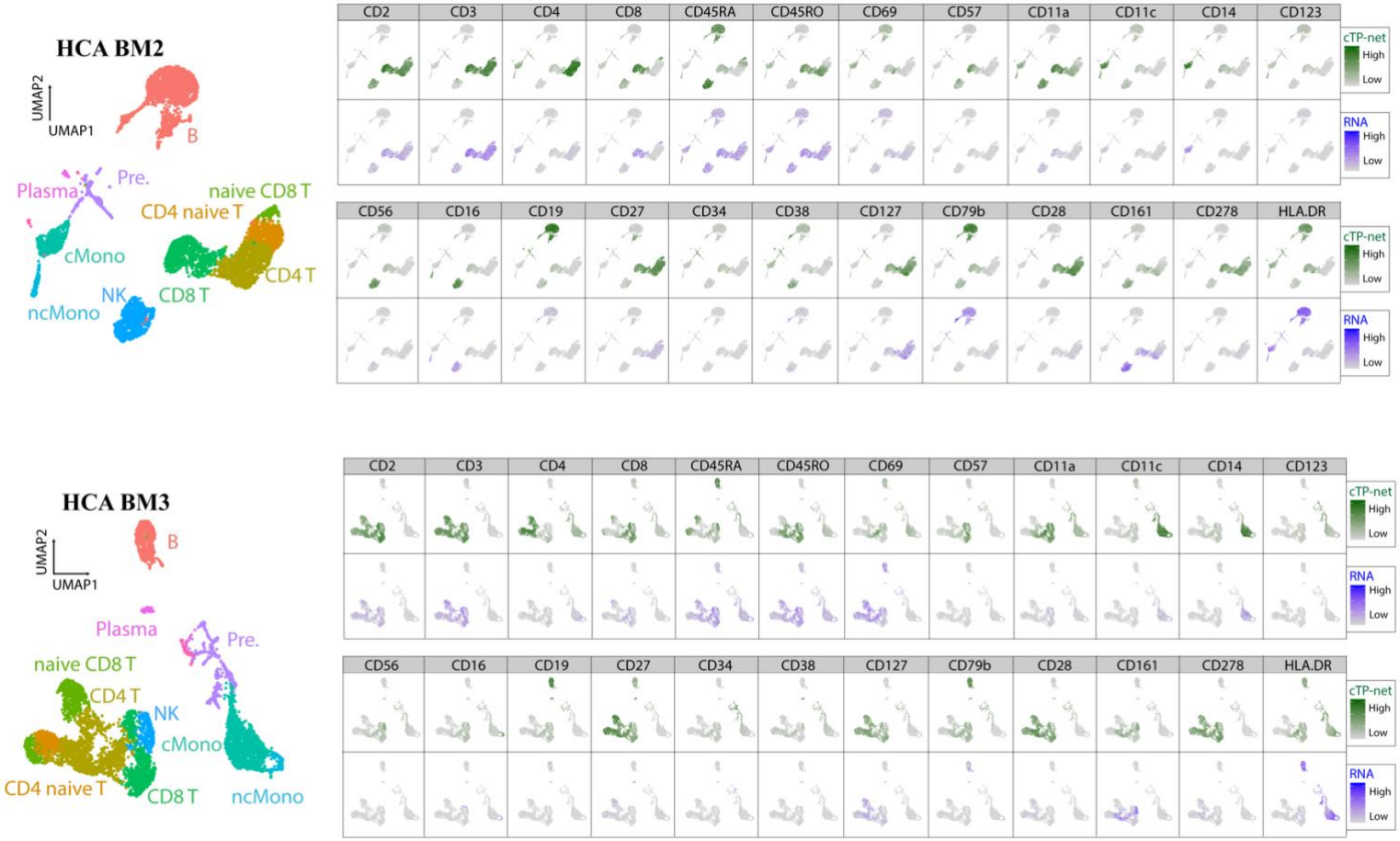
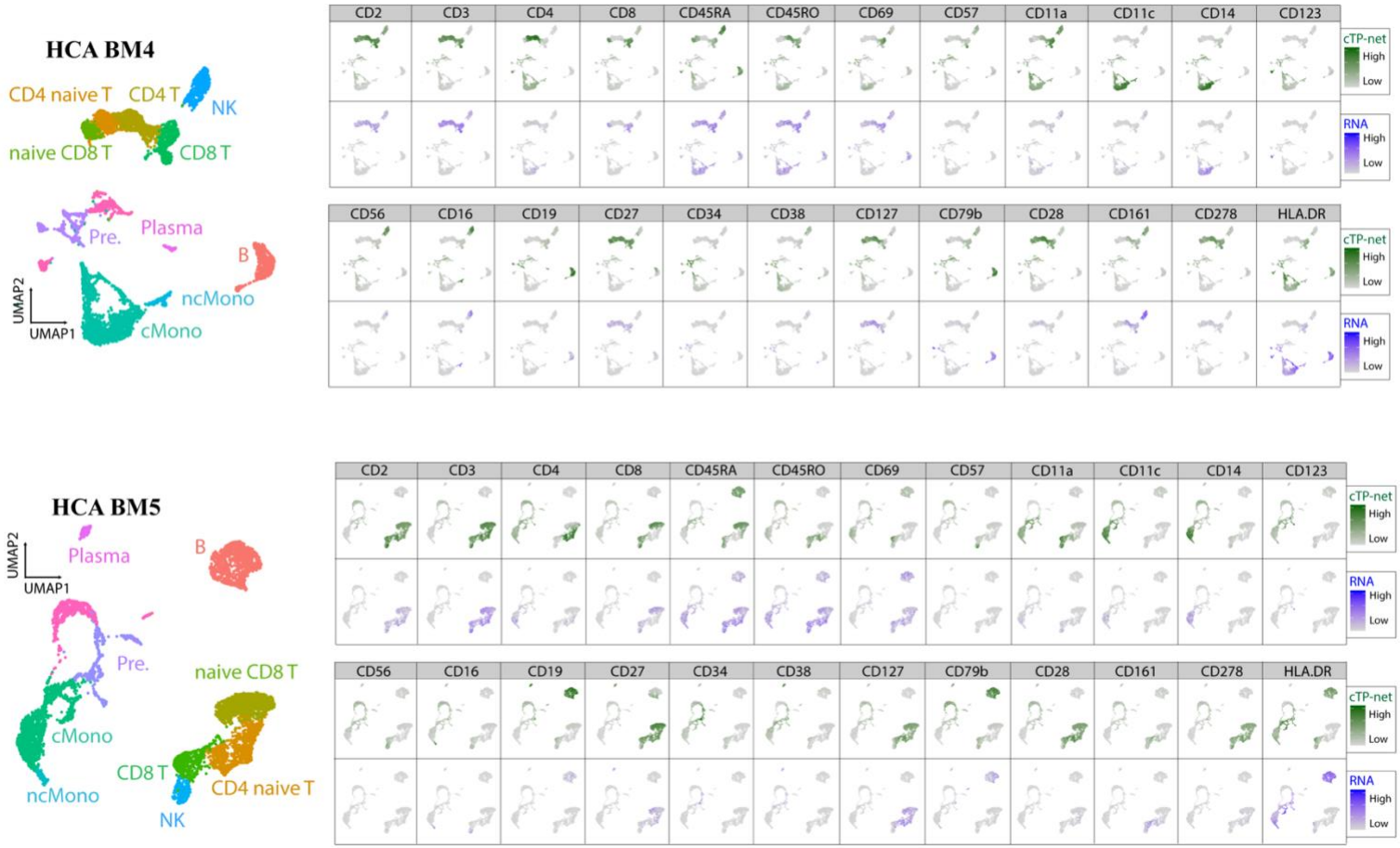
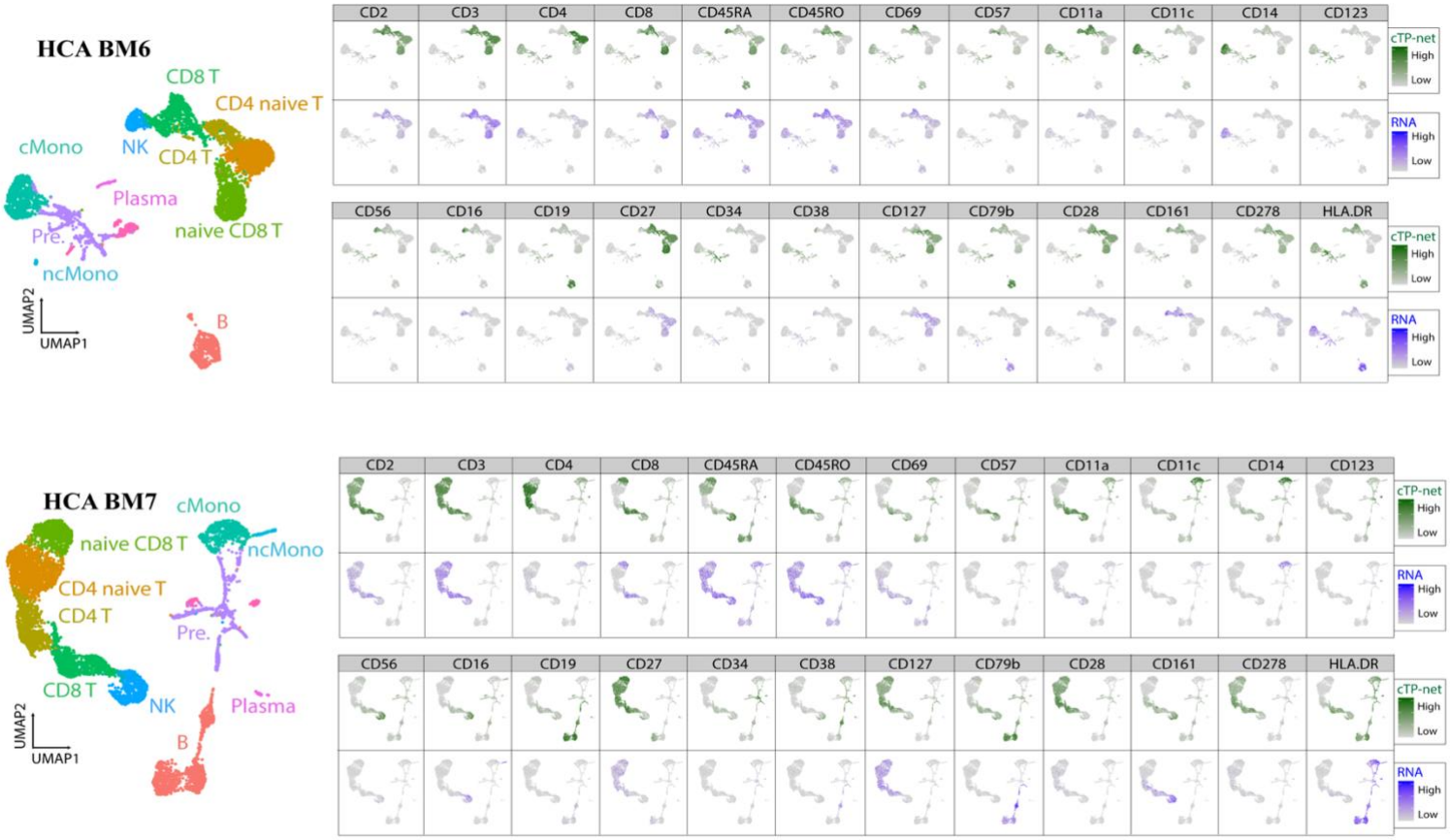

HCA CB8
CD4 T
B
CD8 T
NK
Pre.
cMono
UMAP2
UMAP1
CD2
CD3
CD4
CD8
CD45RA
CD45RO
CD69
CD57
CD11a
CD11c
CD14
CD123
cTP-net
High
Low
RNA
High
Low
CD56
CD16
CD19
CD27
CD34
CD38
CD127
CD79b
CD28
CD161
CD278
HLA.DR
cTP-net
High
Low
RNA
High
Low

Figure 2.10 ***cTP-net prediction on Human Cell Atlas CBMCs by individual.*** *For each individual, we show (1) t-SNE visualization of HCA CBMCs based on expression. B: B cells; CD4 T: CD4 T cells; CD8 T: CD8 T cells; cMono: classic Monocyte; NK: Natural killer cells; Pre.: Precursors. (2) cTP-net imputed protein abundance and RNA of its cognate gene across 24 different surface proteins.*


HCA BM2
UMAP2
UMAP1
B
Pre.
Plasma
naive CD8 T
CD4 naive T
cMono
CD4 T
NK
CD8 T
ncMono
CD2
CD3
CD4
CD8
CD45RA
CD45RO
CD69
CD57
CD11a
CD11c
CD14
CD123
cTP-net
High
Low
RNA
High
Low
CD56
CD16
CD19
CD27
CD34
CD38
CD127
CD79b
CD28
CD161
CD278
HLA.DR
cTP-net
High
Low
RNA
High
Low
HCA BM3
UMAP2
UMAP1
B
Plasma
Pre.
naive CD8 T
CD4 T
NK
cMono
CD4 naive T
CD8 T
ncMono
CD2
CD3
CD4
CD8
CD45RA
CD45RO
CD69
CD57
CD11a
CD11c
CD14
CD123
cTP-net
High
Low
RNA
High
Low
CD56
CD16
CD19
CD27
CD34
CD38
CD127
CD79b
CD28
CD161
CD278
HLA.DR
cTP-net
High
Low
RNA
High
Low


HCA BM4
CD4 naive T
CD4 T
NK
naive CD8 T
CD8 T
Pre.
Plasma
B
ncMono
cMono
UMAP2
UMAP1
CD2
CD3
CD4
CD8
CD45RA
CD45RO
CD69
CD57
CD11a
CD11c
CD14
CD123
CD56
CD16
CD19
CD27
CD34
CD38
CD127
CD79b
CD28
CD161
CD278
HLA.DR
cTP-net
High
Low
RNA
High
Low
HCA BM5
Plasma
B
Pre.
naive CD8 T
cMono
CD8 T
CD4 naive T
ncMono
NK


HCA BM6
CD8 T
CD4 naive T
cMono
NK
CD4 T
Plasma
Pre.
naive CD8 T
ncMono
B
UMAP2
UMAP1
CD2
CD3
CD4
CD8
CD45RA
CD45RO
CD69
CD57
CD11a
CD11c
CD14
CD123
CD56
CD16
CD19
CD27
CD34
CD38
CD127
CD79b
CD28
CD161
CD278
HLA.DR
cTP-net
High
Low
RNA
High
Low

HCA BM7
cMono
naive CD8 T
ncMono
CD4 naive T
CD4 T
Pre.
CD8 T
NK
Plasma
B
UMAP2
UMAP1
CD2
CD3
CD4
CD8
CD45RA
CD45RO
CD69
CD57
CD11a
CD11c
CD14
CD123
CD56
CD16
CD19
CD27
CD34
CD38
CD127
CD79b
CD28
CD161
CD278
HLA.DR
cTP-net
High
Low
RNA
High
Low

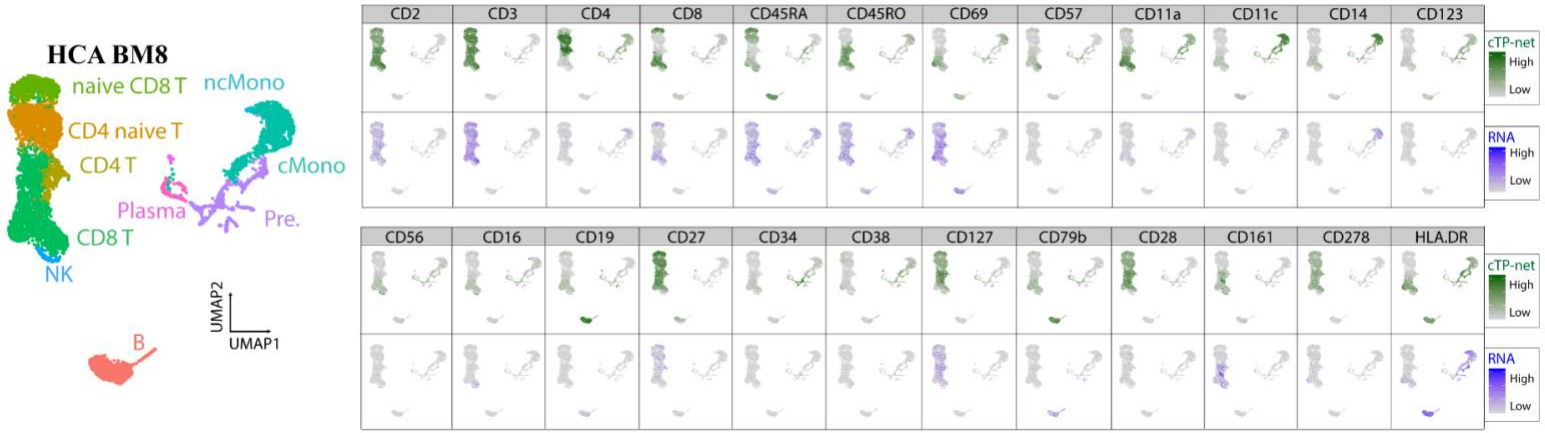

HCA BM8
naive CD8 T
ncMono
CD4 naive T
CD4 T
cMono
Plasma
Pre.
CD8 T
NK
B
UMAP2
UMAP1
CD2
CD3
CD4
CD8
CD45RA
CD45RO
CD69
CD57
CD11a
CD11c
CD14
CD123
cTP-net
High
Low
RNA
High
Low
CD56
CD16
CD19
CD27
CD34
CD38
CD127
CD79b
CD28
CD161
CD278
HLA.DR
cTP-net
High
Low
RNA
High
Low

Figure 2.11 ***cTP-net prediction on Human Cell Atlas BMMCs by individual.*** *For each individual, we show (1) t-SNE visualization of HCA BMMCs based on expression. B: B cells; CD4 T: CD4 T cells; CD8 T: CD8 T cells; Mono: Monocyte; NK: Natural killer cells; Pre.: Precursors. (2) cTP-net imputed protein abundance and RNA of its cognate gene across 12 different surface proteins.*

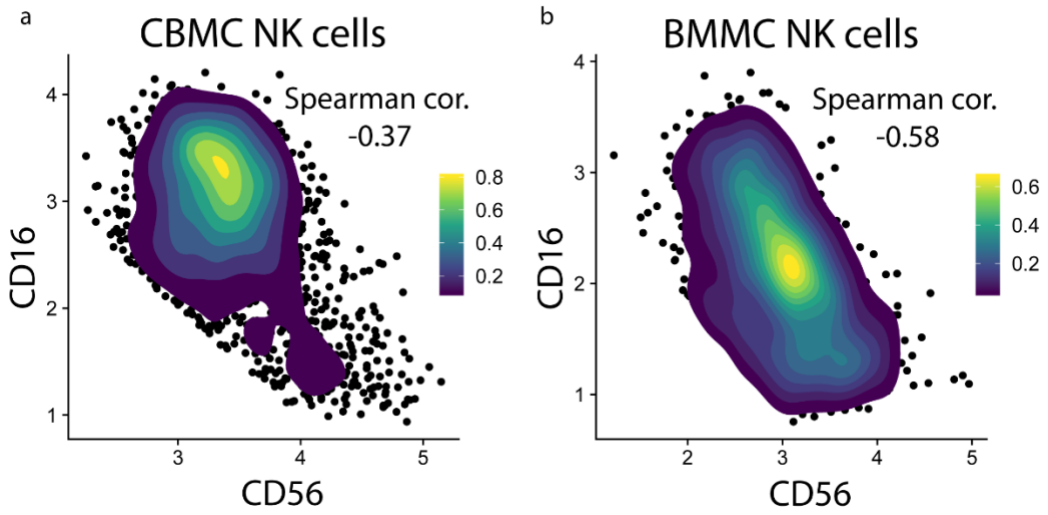


Figure 2.12 ***Contour plot of cells based on imputed CD56 and CD16 abundance in NK cell populations. (a)*** *NK cells across all samples from HCA CBMC.* ***(b)*** *NK cells across all samples from HCA BMMC. Strong negative correlation with two subpopulation observed.*

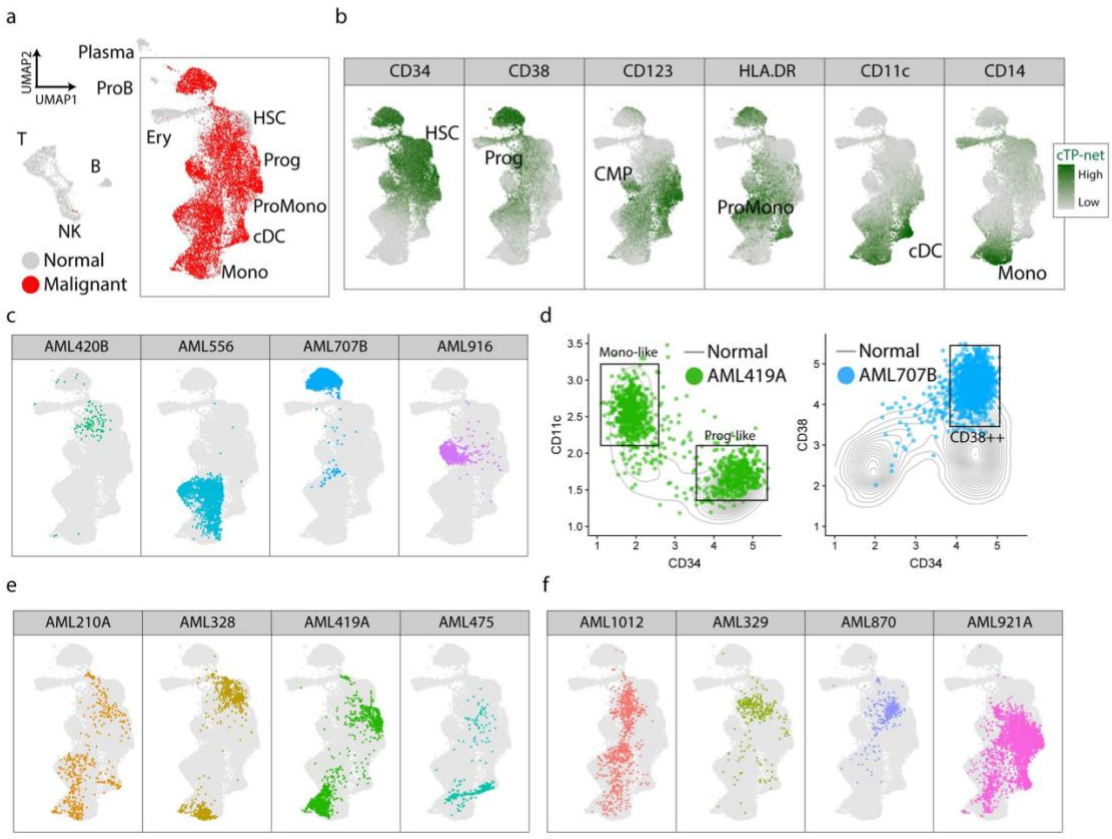


Figure 2.13 ***Imputation results analysis on Acute Myeloid Leukemia data sets. (a)*** *UMAP visualization of normal cells and malignant cells from 12 AML samples at Day0 based on imputed protein abundance (red: malignant cells; grey: normal cells).* ***(b)*** *UMAP visualization of the myeloid trajectory. cTP-net imputed protein abundance of markers that perfectly recapitulate the myeloid development.* ***(c, e, f)*** *UMAP visualization of the myeloid trajectory with corresponding malignant cells from AML sample highlighted.* ***(d)*** *Plot of normal cells (grey contour) and AML malignant cells (dots) based on imputed protein expression.*

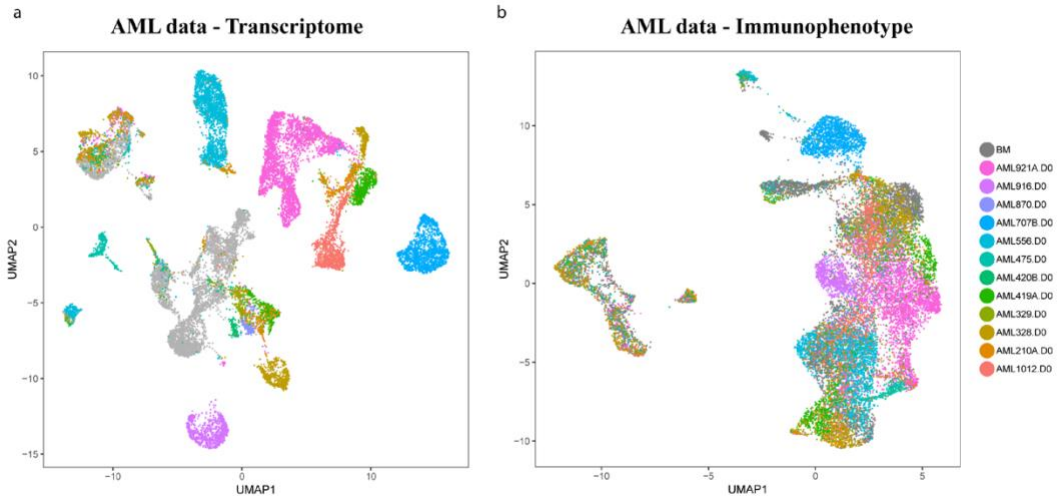


Figure 2.14 ***UMAP plots of AML data set, colored by samples. (a)*** *Dimension reduction on transcriptome (RNAs).* ***(b)*** *Dimension reduction on imputed surface proteins.*

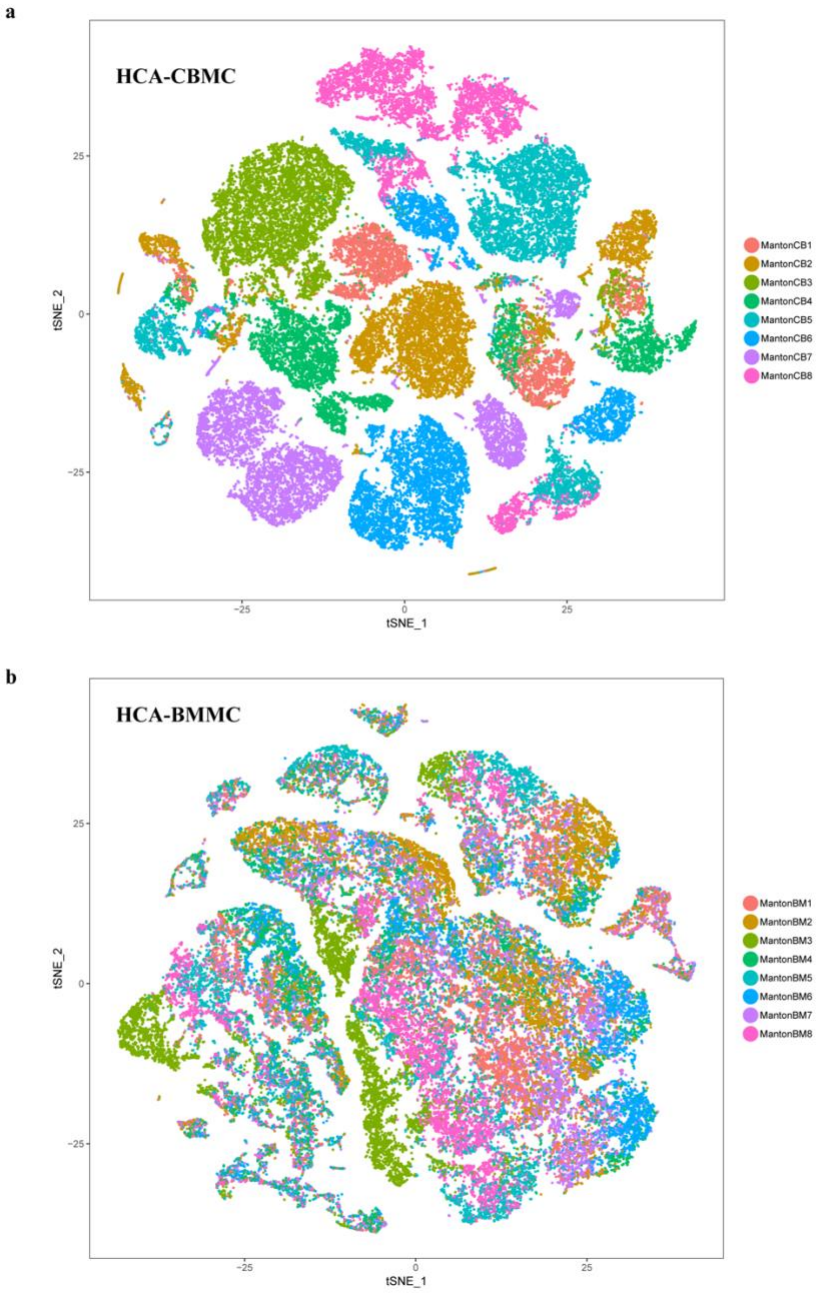


Figure 2.15 ***Human Cell Atlas t-SNE plot based on normalized expression. (a)*** *t-SNE plot on Human Cell Atlas CBMCs based on normalized expression. Color indicates sample IDs.* ***(b)*** *t-SNE plot on Human Cell Atlas BMMCs based on normalized expression. Color indicates sample IDs. Strong batch effects observed in both data sets.*

Table 2.1 Summary table of five data sets analyzed in this study

Data	Technology	Cell population	# of subjects	# of cells	# of genes	# of proteins	# of cell types
CITE-PBMC	CITE-seq	PBMC	1	7667	13517	10	8
CITE-CBMC	CITE-seq	CBMC	1	8005	14505	10	12
REAP-PBMC	REAP-seq	PBMC	1	4326	10811	10	NA
CITE-BMMC	CITE-seq	BMMC	1	33455	17009	25	NA
HCA-CBMC	10x	CBMC	8	260,000	12611	NA	NA
HCA-BMMC	10x	BMMC	8	270,000	12611	NA	NA

Table 2.2 Cell type summary of CITE-seq data sets

Data	Cell types
CITE-PBMC	B, CD8 T-1, CD4 T, NK, DC, CD14+CD16+ Mono, CD14-CD16+ Mono, CD8 T 2
CITE-CBMC	B, CD8 T, CD4 T, NK, DC, CD14+ Mono, CD16+ Mono, pDC, CD34+, Eryth, Unknown

Table 2.3 Top 20 highest influence score genes for each protein in CITE-PBMC data set

CD3	CD4	CD8	CD2	CD45RA	CD57	CD16	CD14	CD11c	CD19
CD3D	*CD8B*	*CD8B*	*CCL5*	*KLRB1*	*NUDT6*	*CHL1*	*C1orf115*	*CFD*	*CCL5*
IL7R	*CD8A*	*CD8A*	*IL7R*	*CCL5*	*MZT2A*	*RP11-242C19.2*	*PEAK1*	*MAL*	*CD8B*
CD8B	*RP11-291B21.2*	*CCL5*	*RP4-539M6.22*	*EIF1AX*	*ATP2A2*	*GCSH*	*ALDH7A1*	*ANKRD36C*	*RN7SL600P*
FCER1G	*CCL5*	*TRDC*	*LTV1*	*CD7*	*IQCE*	*NRL*	*CYBB*	*BLOC1S3*	*MYO1D*
TRDC	*NCR3*	*RP11-291B21.2*	*RP11-452L6.5*	*TST*	*PKNOX1*	*DBF4*	*ISYNA1*	*IGLL5*	*HSF2*
AKR7A2	*KLRB1*	*CHMP7*	*FBXO10*	*MFSD7*	*FBXW8*	*SPHK2*	*LMAN1*	*RP11-159G9.5*	*AC142528.1*
HELLS	*DDIT3*	*ZAP70*	*LINC00384*	*ZFAS1*	*CTA-217C2.1*	*CDKL1*	*FAM162A*	*ARMCX1*	*DNAJA3*
ALG10	*CTD-2547L16.1*	*BMP8B*	*ACAP2*	*TAPSAR1*	*CNOT11*	*TIMM21*	*SLC4A7*	*SLC6A16*	*GLB1L*
FGD5-AS1	*C18orf25*	*FAH*	*PPCDC*	*PLEKHF1*	*HSD17B4*	*MRPS18C*	*MIER3*	*LRRC16A*	*LIMD2*
COMMD7	*FPGT*	*AC009299.3*	*ANKRD39*	*CTBP1-AS2*	*CLEC4E*	*C7orf43*	*SLC11A2*	*TRAF1*	*DTX3L*
CTA-292E10.8	*NETO2*	*CMKLR1*	*AIM2*	*CYP27A1*	*FAM98C*	*GORASP2*	*ZAP70*	*PABPN1*	*PTCD2*
ZC2HC1A	*GDAP1*	*ENTPD1*	*TTLL12*	*MAN1A2*	*PRMT1*	*CTD-2555C10.3*	*MAP4*	*ADM*	*LPAR1*
INADL	*CSTF1*	*PIK3CA*	*GABBR1*	*FAM115C*	*SLC25A11*	*LEPROT*	*TTTY15*	*KIAA0319L*	*ZNF649*
SHISA4	*RP11-159H10.3*	*WDR7*	*DCUN1D4*	*CST3*	*TCEANC2*	*RUSC1*	*HS1BP3*	*MRPL4*	*HLA-DRB5*
DCAF4	*RP11-451M19.3*	*HEG1*	*CPD*	*NAIF1*	*LCTL*	*POLR2L*	*PRPSAP1*	*NDRG1*	*LIN54*
HPGDS	*ENTPD1-AS1*	*NPAT*	*RAPGEFL1*	*RP11-83N9.5*	*CAPN1*	*RP11-85A1.3*	*ZBTB38*	*FAM63A*	*USP32*
PACSIN1	*SLC4A10*	*7-Sep*	*U91328.20*	*FCGR3A*	*VPS26A*	*FKBP7*	*PIK3R1*	*RPL34*	*AIM2*
ARID4B	*FAAH2*	*CDT1*	*EIF4H*	*CCDC163P*	*ECHS1*	*RNF24*	*PIGG*	*FAM118B*	*SLC12A7*
ATP11A	*AP5B1*	*QRICH1*	*AC073115.7*	*POLR1C*	*FLVCR1-AS1*	*TBXAS1*	*DESI2*	*UBQLN4*	*ZNF671*
RN7SL521P	*DHPS*	*AP2M1*	*RP11-4O1.2*	*PRSS35*	*RP11-421L21.2*	*WDR83*	*SIRT5*	*FKBP15*	*TNNI2*

Table 2.4 Summary table of different cTP-net models

Differences to the finalized model	Correlation
Without SAVER-X denoising, without MB structure	0.961±0.0004
Without MB structure	0.968±0.0005
Without SAVER-X denoising	0.959 ±0.0005
L2 loss	0.969±0.0002
Set bottle neck layer to 256 nodes (128 in final model)	0.968±0.0003
Set bottle neck layer to 64 nodes (128 in final model)	0.968±0.0003
With additional shared layers	0.969±0.0004
With SeLU activation function	0.966±0.0002
With Dropout layer between layer1 and layer2	0.966±0.001
Exclude genes corresponding to targeted proteins	0.967±0.0001
Final model	**0.970**±0.0003

Table 2.5 List of surface proteins and corresponding genes

Surface protein	Corresponding gene
CD3	*CD3D,CD3E,CD3G,CD247*
CD4	*CD4*
CD8	*CD8A,CD8B*
CD45RA	*PTPRC*
CD56	*NCAM1*
CD2	*CD2*
CD16	*FCGR3A*
CD11c	*ITGAX*
CD14	*CD14*
CD19	*CD19*
CD34	*CD34*
CD57	*B3GAT1*
CD11a	*ITGAL*
CD123	*IL3RA*
CD127	*IL7R*
CD161	*KLRB1*
CD27	*CD27*
CD278	*ICOS*
CD28	*CD28*
CD38	*CD38*
CD45RO	*PTPRC*
CD69	*CD69*
CD79b	*CD79B*
HLR.DR	*HLA-DRA,HLA-DRB1,HLA-DRB5*

Table 2.6 Gene set enrichment analysis on cell-immunophenotype pairs that cTP-net predict well in CITE-PBMC data set

Surface protein	Cell type	GO pathways
CD45RA	**CD14-CD16+Mono**	GO_CATABOLIC_PROCESS
		GO_PROTEIN_LOCALIZATION
		GO_REGULATION_OF_CELLULAR_COMPONENT_BIOGENESIS
		GO_CELLULAR_RESPONSE_TO_STRESS
		GO_CELLULAR_RESPONSE_TO_DNA_DAMAGE_STIMULUS
		GO_RNA_BINDING
		GO_ESTABLISHMENT_OF_LOCALIZATION_IN_CELL
		GO_CELL_CYCLE
		GO_SINGLE_ORGANISM_BIOSYNTHETIC_PROCESS
		GO_CELLULAR_MACROMOLECULE_LOCALIZATION
CD11c	**CD14-CD16+Mono**	GO_CELLULAR_RESPONSE_TO_STRESS
		GO_NEGATIVE_REGULATION_OF_GENE_EXPRESSION
		GO_POSITIVE_REGULATION_OF_BIOSYNTHETIC_PROCESS
		GO_POSITIVE_REGULATION_OF_GENE_EXPRESSION
		GO_CELL_CYCLE
		GO_POSITIVE_REGULATION_OF_PROTEIN_METABOLIC_PROCESS
		GO_NEGATIVE_REGULATION_OF_NITROGEN_COMPOUND_METABOLIC_PROCESS
		GO_CYTOSKELETON
		GO_CHROMOSOME
		GO_ENZYME_BINDING
CD45RA	**CD8 T 2**	GO_ENZYME_BINDING
		GO_RNA_BINDING
		GO_RIBONUCLEOPROTEIN_COMPLEX
		GO_REGULATION_OF_TRANSCRIPTION_FROM_RNA_POLYMERASE_II_PROMOTER

		GO_CELL_CYCLE
		GO_RNA_PROCESSING
		GO_POSITIVE_REGULATION_OF_BIOSYNTHETIC_PROCESS
		GO_CYTOSKELETON
		GO_RIBONUCLEOTIDE_BINDING
		GO_POSITIVE_REGULATION_OF_GENE_EXPRESSION
CD45RA	**CD4 T**	GO_REGULATION_OF_IMMUNE_SYSTEM_PROCESS
		GO_IMMUNE_SYSTEM_PROCESS
		GO_VACUOLE
		GO_SMALL_MOLECULE_METABOLIC_PROCESS
		GO_ORGANONITROGEN_COMPOUND_METABOLIC_PROCESS
		GO_ESTABLISHMENT_OF_LOCALIZATION_IN_CELL
		GO_POSITIVE_REGULATION_OF_MULTICELLULAR_ORGANISMAL_PROCESS
		GO_ENDOPLASMIC_RETICULUM
		GO_REGULATION_OF_TRANSCRIPTION_FROM_RNA_POLYMERASE_II_PROMOTER
		GO_PROTEIN_LOCALIZATION
CD11c	**CD14+CD16+ Mono**	GO_POSITIVE_REGULATION_OF_GENE_EXPRESSION
		GO_DNA_REPLICATION
		GO_POSITIVE_REGULATION_OF_MOLECULAR_FUNCTION
		GO_POSITIVE_REGULATION_OF_BIOSYNTHETIC_PROCESS
		GO_SINGLE_ORGANISM_BIOSYNTHETIC_PROCESS
		GO_DNA_DEPENDENT_DNA_REPLICATION
		GO_PHOSPHATE_CONTAINING_COMPOUND_METABOLIC_PROCESS
		GO_CELL_JUNCTION
		GO_CYTOKINE_RECEPTOR_BINDING
		GO_ORGANONITROGEN_COMPOUND_BIOSYNTHETIC_PROCESS
CD45RA	**DC**	GO_NEGATIVE_REGULATION_OF_NITROGEN_COMPOUND_METABOLIC_PROCESS

		GO_POLY_A_RNA_BINDING
		GO_CHROMOSOME_ORGANIZATION
		GO_REGULATION_OF_DNA_METABOLIC_PROCESS
		GO_RNA_BINDING
		GO_MACROMOLECULAR_COMPLEX_BINDING
		GO_PHOSPHATE_CONTAINING_COMPOUND_METABOLIC_PROCESS
		GO_NEGATIVE_REGULATION_OF_GENE_EXPRESSION
		GO_ESTABLISHMENT_OF_LOCALIZATION_IN_CELL
		GO_DNA_METABOLIC_PROCESS
CD11c	**DC**	GO_ENZYME_BINDING
		GO_RIBONUCLEOTIDE_BINDING
		GO_ESTABLISHMENT_OF_LOCALIZATION_IN_CELL
		GO_NEGATIVE_REGULATION_OF_PROTEIN_METABOLIC_PROCESS
		GO_IMMUNE_SYSTEM_PROCESS
		GO_ORGANONITROGEN_COMPOUND_BIOSYNTHETIC_PROCESS
		GO_PHOSPHATE_CONTAINING_COMPOUND_METABOLIC_PROCESS
		GO_PHOSPHORYLATION
		GO_NEGATIVE_REGULATION_OF_PROTEIN_MODIFICATION_PROCESS
		GO_TRANSFERASE_ACTIVITY_TRANSFERRING_PHOSPHORUS_CONTAINING_GROUPS

CHAPTER 3 INTEGRATIVE DNA COPY NUMBER DETECTION AND GENOTYPING FROM SEQUENCING AND ARRAY-BASED PLATFORMS WITH PENN MEDICINE BIOBANK

3.1 Introduction

Copy number variations (CNV) are large chunks of DNA that have been deleted or duplicated during evolution, leading to polymorphisms in their numbers of copies in the observed population. Studies have shown that CNV is an important type of variation in the human genome, some of which playing key roles in disease susceptibility [130-132]. Accurate identification and genotyping of CNV is important for population genetic and disease studies, and can lead to improved understanding of disease mechanisms and discovery of drug targets [133-135]. To profile CNV, earlier studies relied on array-based technologies such as array comparative genome hybridization (CGH) or single-nucleotide polymorphism (SNP) genotyping arrays, while in recent years, next generation sequencing (NGS) technologies have allowed for high resolution CNV profiling [136-143]. With the drop in sequencing cost, many large cohort profile both array data and NGS data from same sample. Such design allows better sensitivity and specificity of CNV detection. We recently developed a statistical framework, integrated Copy Number Variation caller (iCNV), that can be applied to study design of combination of SNP and sequencing data [144]. Compared to existing approaches, iCNV improves copy number detection accuracy in three ways: (1) utilization of B allele frequency information from sequencing data, (2) integration of sample matched SNP-array data, and (3) integration of improved platform-specific normalization for sequencing coverage. iCNV produces a cross-platform joint segmentation of each sample's genome into deleted, duplicated, and normal regions, and further infers integer copy numbers in deletion and duplication regions.

Recent years' developments of large genomic biobank propose great opportunity for CNV studies across many phenotypes[145, 146]. The Penn Medicine BioBank (PMBB), a diverse

cohort, currently consists of paired SNP array and whole exome sequencing (WES) data from 2219 African ancestry samples and 8078 European ancestry samples. A complete profile of CNVs of all PMBB samples in companion with detailed patient health information can provide a great resources for researchers to understand the relationship between germline CNVs and various phenotype. In order to adjust to large number of samples, we improve iCNV with an efficient Map-Reduce algorithm for CNV detection that reduce computation time and boost robustness [147].

3.2 Methods

3.2.1 Penn Medicine BioBank

PMBB recruits participants by enrolling at the time of appointment through the University of Pennsylvania Health System. Patients are asked to donate either blood or a tissue sample and allow researchers access to their electronic health record (EHR). This provides researchers with access to a large resources of genomic data with attached health information. PMBB currently consists of 8078 European ancestry samples and 2219 African ancestry samples with paired SNP array and WES data.

3.2.2 Pipeline overview

Fig. 3.1 shows an overview of iCNV analysis pipeline. Input data depends on experiment design: When both SNP array and NGS data are available, the input includes (i) SNP log R ratio (LRR) and (ii) B allele frequency (BAF), which quantify, respectively, relative probe intensity and allele proportion, and (iii) sequencing mapped reads (BAM file) [146, 148]. For sequencing data, iCNV also receives target positions (BED file) for read depth background normalization. In WES, the targets are exons, while for WGS, iCNV automatically bins the genome and treats each bin as a target (the default bin size is 1kb). iCNV first performs cross-sample bias correction for sequencing data using CODEX and computes a Poisson log-likelihood ratio (PLR) for each target [137]. As suggested, samples with different ethnicity needed to be separated for analysis. In

addition, the sequencing batch information of the samples is unavailable. In order to have an unbiased normalization method, we performed permutation-based test which will introduce in the next section (Fig. 3.2). Heterozygous SNPs are detected and BAFs are computed within target regions using SAMTOOLS [148]. Integrated CNV detection is then conducted through a hidden Markov model (HMM) that treats the array intensity, array BAF, sequencing PLR and sequencing BAF as observed emissions from a hidden copy number state. The HMM segments the genome of each sample into regions of homogeneous copy number and outputs an integrated Z-score for each position that summarizes the evidence for an abnormal copy number at that position. Integer-valued copy numbers are then estimated in regions of high absolute Z-score, utilizing information from all platforms. Finally, we filters out small CNVs with size less than 10kb as well as untrustful regions, such as immunoglobulin regions.

3.2.3 Map-Reduce framework for efficient and robust CNV detection

Due to large number of samples and missing batch information, we design a map-reduce framework aiming to reduce computational time and improve CNV detection robustness. Analysis shows that the step of calculating Poisson log likelihood ratio is the bottleneck steps. This is due to large samples size, intractable RAM, multi-core inability as well as unavailable of batch information. As a result, we randomly partition the samples into batches of size around 100 and remove the biases at batch level illustrated in Fig 3.2. In this computational step, we map the data set into a number of workers in the computer cluster, where the normalization was performed per worker (i.e. the map step). We further combine the normalized data in individual batch into a full dataset and apply HMM algorithm for CNV detection (i.e. the reduce step). Owning to the fact that we do not have batch information, we permute the batch assignment 5 times and take a majority vote of the CNV calls to ensure detection robustness. Such framework reduces the computational time by 100 folds and allows higher confidence in CNV calls without prior batch information.

3.3 Results

3.3.1 CNV summary of samples

Fig. 3.3 provides an example of heatmap of the CNV scores across 120 samples, with blue illustrates higher chance of duplication and red illustrates higher chance of deletion. Dark blue dots and dark red dots indicates CNV calls of duplication and deletion respectively. The CNV distribution of European ancestry (EUR) samples is illustrated in Fig. 3.4a. iCNV detects on average 34.1 deletions and 11.3 duplications per EUR sample. Fig. 3.4c shows the CNV distribution of African ancestry (AFR) samples. iCNV detects on average 38 deletions and 10.6 duplications per AFR sample, with trend similar to EUR samples. However, as we noticed, there are clearly higher number and bigger size of homozygous deletions and duplications detected in AFR than EUR (Fig. 3.4b, d). This might be due to the fact that the WES data was mapped to a human genome reference with majority of reference samples from European ancestry. The high number of homozygous deletion and duplication in the AFR might just be gaps and diversities that was not captured in the reference genome. However, further investigation of the CNV burden differences between AFR samples and EUR samples are necessary.

3.3.2 Comparison with CLAMMS

The PMBB samples have been applied to a computational method called Copy number estimation using Lattice-Aligned Mixture Models (CLAMMS), which utilize only the WES read depth information for CNV detection [149]. On average, iCNV identified more and bigger CNV cases comparing the CLAMMS, which is contributed by integration of both allele frequency information and additional resources of SNP array. Fig. 3.5 shows an example of 1Mb regions of *TG* gene where iCNV detect CNVs but CLAMMS do not. Sample UPENN6848 and sample UPENN10001043 both show that the deletion regions are covered by only few exons but many SNPs, thus iCNV provides additional sensitivity as it adopts SNP information (Fig. 3.5bc). Another example is 800kb region of gene *RIMS2* (Fig. 3.6). For sample UPENN4733, even though both

CLAMMS and iCNV detected this duplication, iCNV provides higher resolution in terms of the segmentation point with SNP array information (Fig. 3.6b). Sample UPENN10010167 is an another example of duplication regions that covered by only few exons but many SNPs (Fig. 3.6c). Actually, as shown in the iCNV paper, we find that an integrated analysis yields more deletion and duplications than single platforms. More importantly, when comparing the integrated analysis with a simple intersection or union of results from a separate analysis of each individual platform, iCNV achieves specificity close to intersection and sensitivity of the union (Fig. 3.7). A signal that is moderate in both platforms would be present in the integrated call set but not in the union call set. A signal that is only present in one platform but absent in the other would be present in the union call set but not detected during integration. Compared to taking a simple union, combining the two platforms improves resolution, thus improving CNV detection power, and integration by the hidden Markov model allows one platform to “check” the calls of the other, thus improving robustness.

3.4 Conclusion

We have detected the CNV profile across 10297 samples in the PMBB with both SNP-array and WES data using iCNV. Comparing with method that only utilizes WES read depth features, iCNV shows higher sensitivity and robustness. In addition, through a Map-Reduce framework with permutation, we reduce the total computation time by 100 folds and allow robust normalization step. This work provides an rich resources for understanding CNVs and pave the ways to many potential studies of PMBB such as CNV risk score [145], PheWAS analysis [150] and CNV variation between ethnicities.

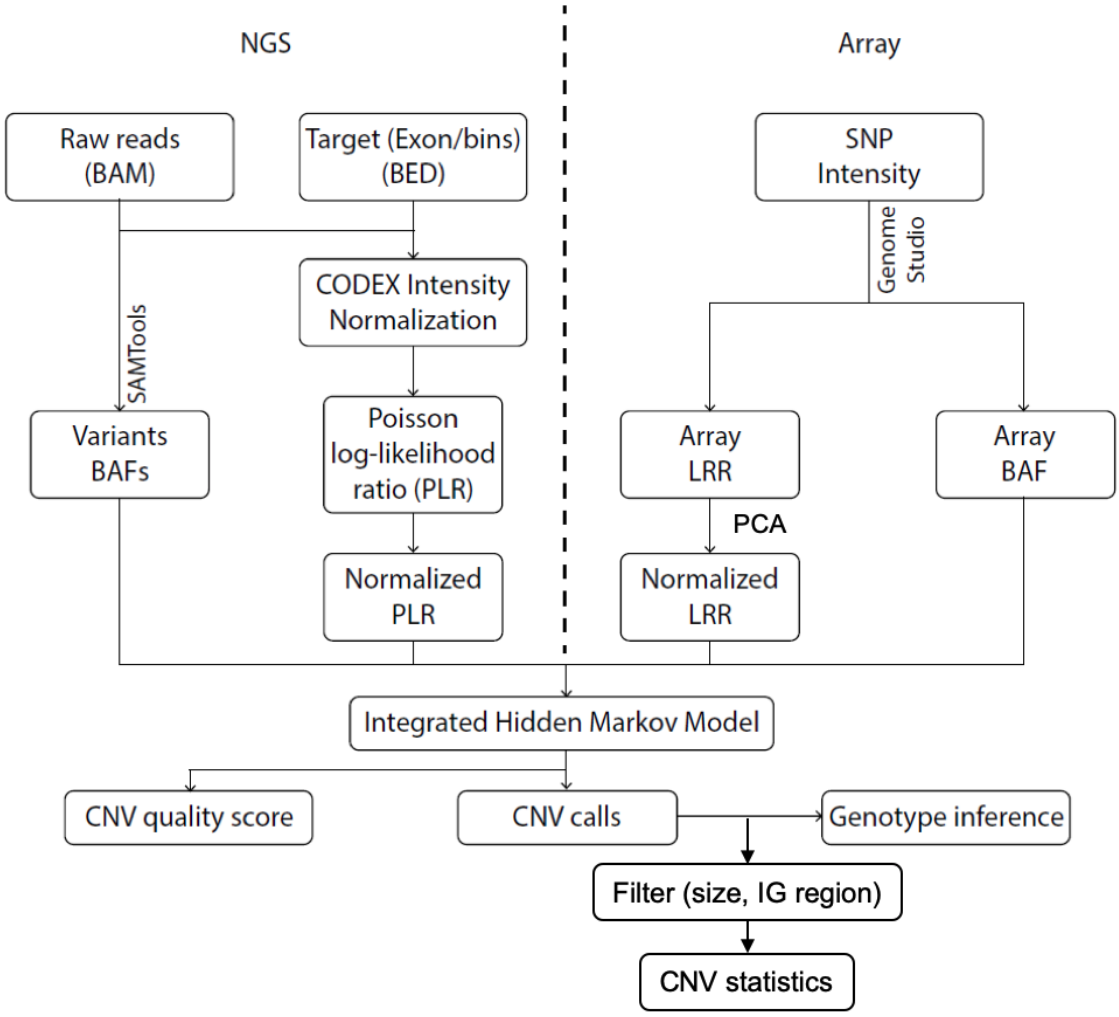


Figure 3.1 ***iCNV analysis pipeline including data normalization, CNV calling and genotyping using NGS and array data.*** *For NGS data, the first step is to normalize coverage using CODEX and calculate a Poisson log-likelihood ratio (PLR), further converted to a normalized LRR by a z-transformation. The heterozygous single nucleotide positions are then found and BAF computed using SAMTools. For array data, we obtain log R ratios and BAF from raw SNP intensity data, then normalize the log R ratios. The integrated Hidden Markov Model takes these inputs and generate integrated CNV calls with quality scores. Finally, genotypes are inferred for each CNV region.*

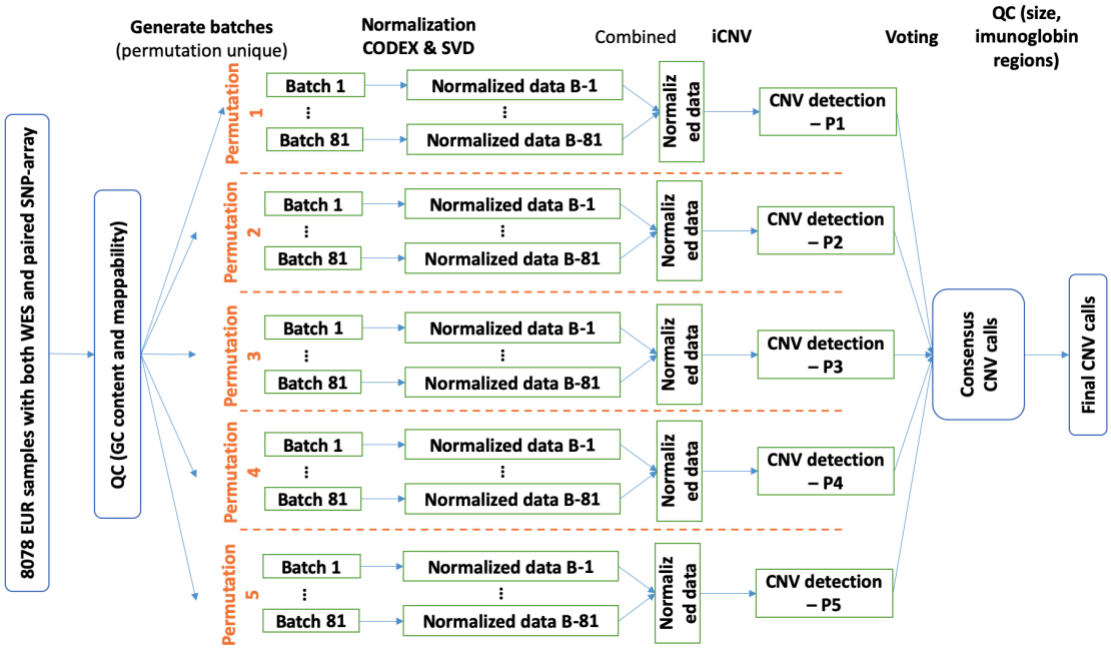


Figure 3.2 ***Map-reduce framework for CNV profiling of PMBB data set.*** *Here, we select the pipeline for 8078 EUR samples for illustration.*

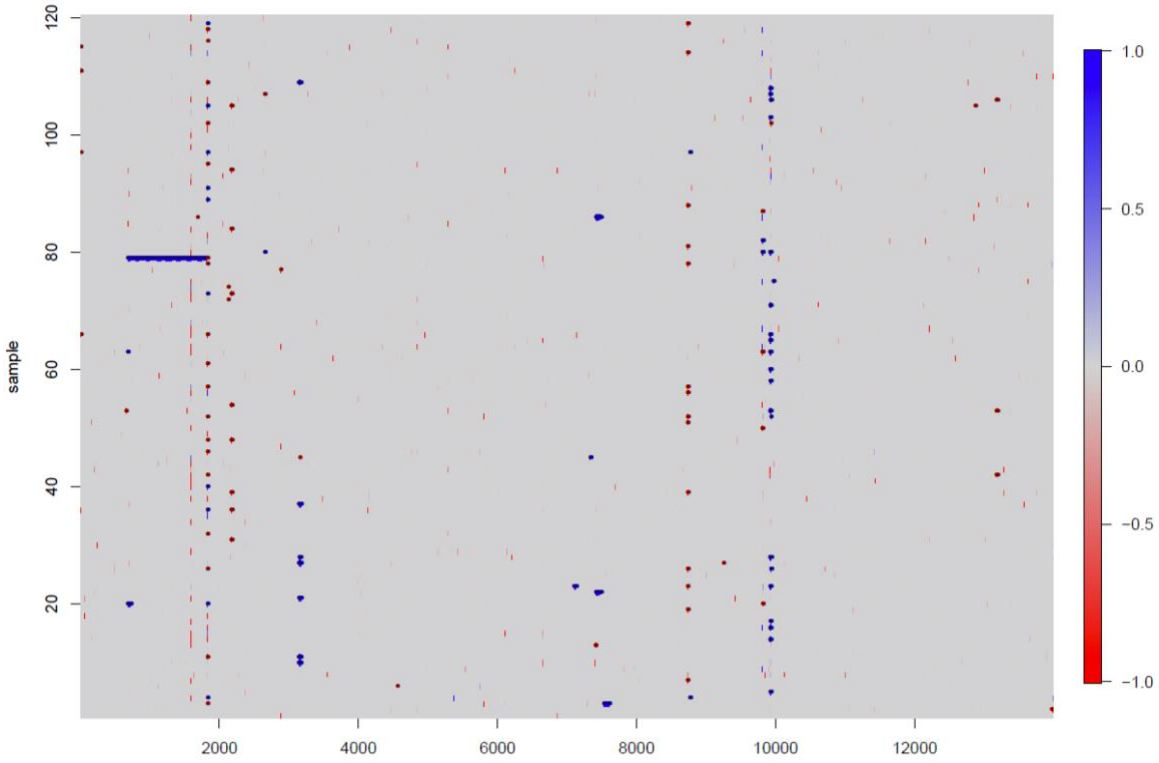


Figure 3.3 ***CNV detection by iCNV (120 example individual chr22, CNV>10kb).*** *Heat map indicates CNV scores (blue indicates more likely to be duplication and red indicates more likely to be deletion) and CNV calling (dark blue dots: duplication; dark red dots: deletion). Here, each row represent a sample and each column represent a hidden state.*

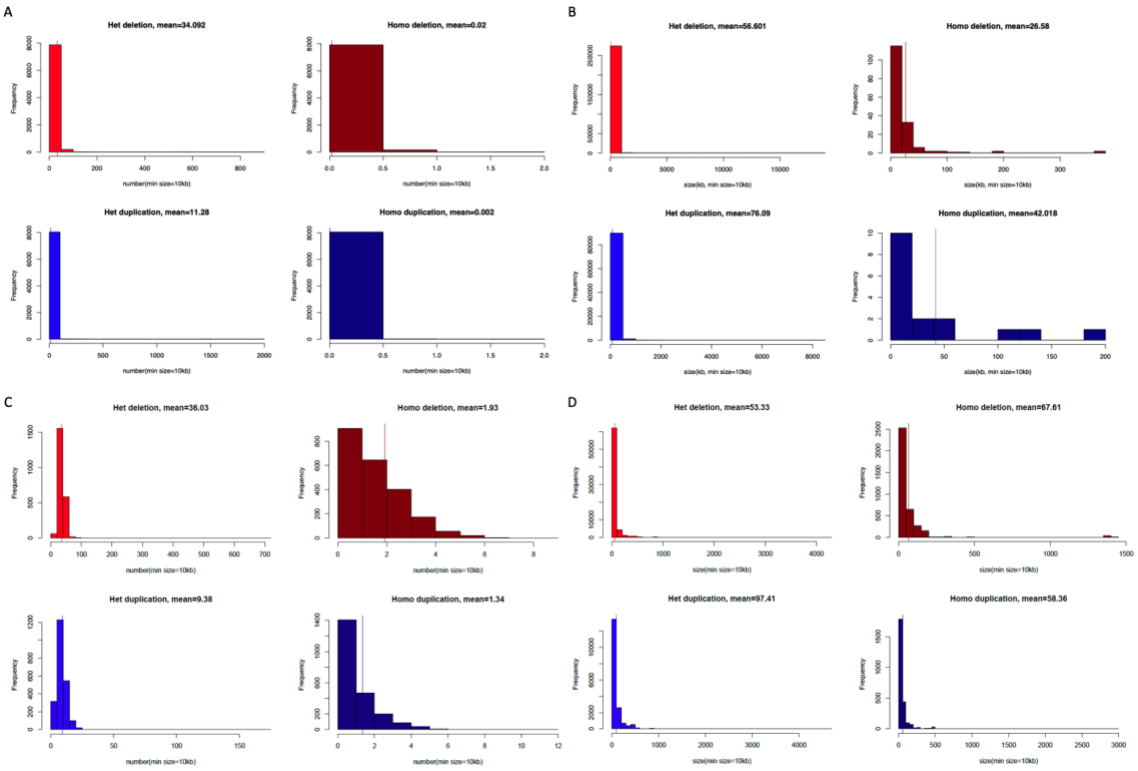


Figure 3.4 ***Summary statistics of iCNV results. a*** *Distribution of number CNVs per sample across 8087 EUR samples.* ***b*** *Distribution of size of CNVs across 8087 EUR samples.* ***c*** *Distribution of number CNVs per sample across 2219 AFR samples.* ***d*** *Distribution of size of CNVs across 2219 AFR samples.*

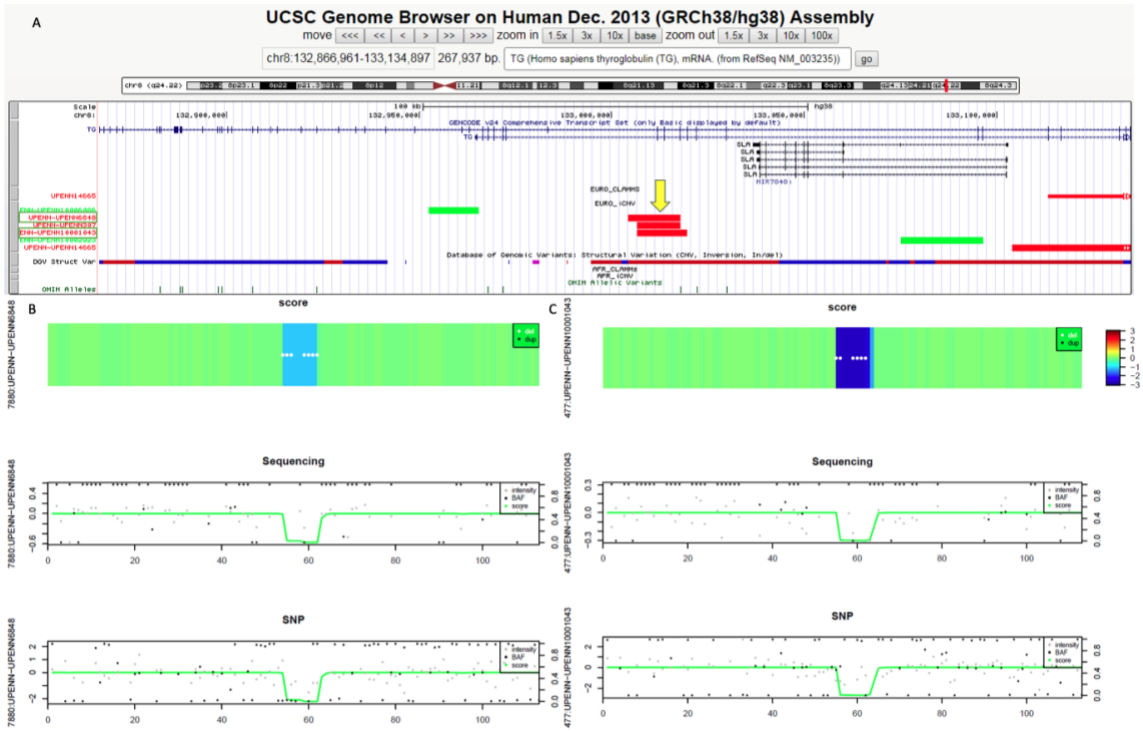


Figure 3.5 ***iCNV vs. CLAMMS of 1Mb region around gene TG. a*** *UCSC Genome Browser shows the CNV calling result at this regions of CLAMMS and iCNV. Here, red bar indicates tentative deletion and green bar indicating tentative duplication. Yellow arrow indicates regions of focus for* ***b*** *and* ***c****.* ***b, c*** *iCNV plot. First panel shows the iCNV score heatmap, with white dots indicating deletion detected. Second and third panel show normalized data distribution of sequencing and SNP. Grey dots indicate intensity, black dots indicate BAF and green line shows iCNV score.*

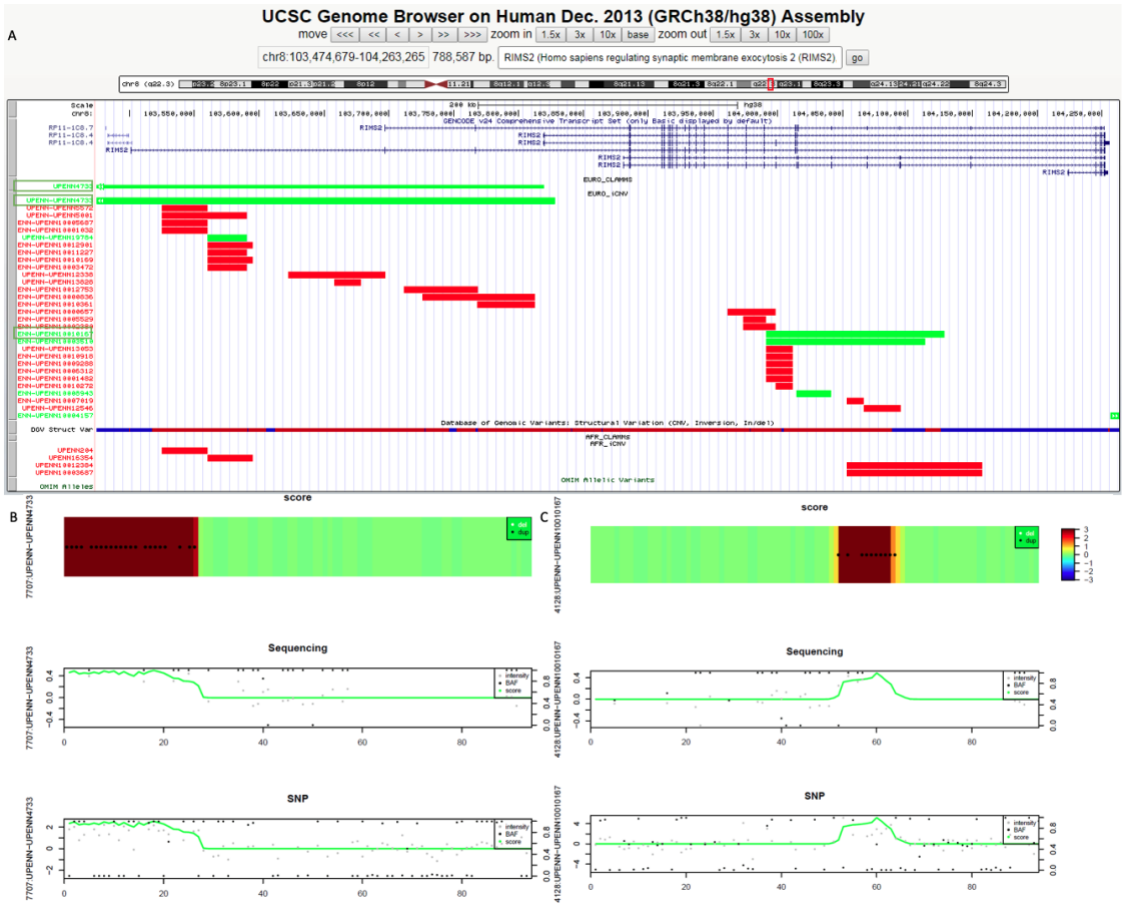


Figure 3.6 ***iCNV vs. CLAMMS of 800kb region around gene RIMS2. a*** *UCSC Genome Browser shows the CNV calling result at this regions of CLAMMS and iCNV. Here, red bar indicates tentative deletion and green bar indicating tentative duplication.* ***b, c*** *iCNV plot. First panel shows the iCNV score heatmap, with black dots indicating duplication detected. Second and third panel show normalized data distribution of sequencing and SNP. Grey dots indicate intensity, black dots indicate BAF and green line shows iCNV score.*

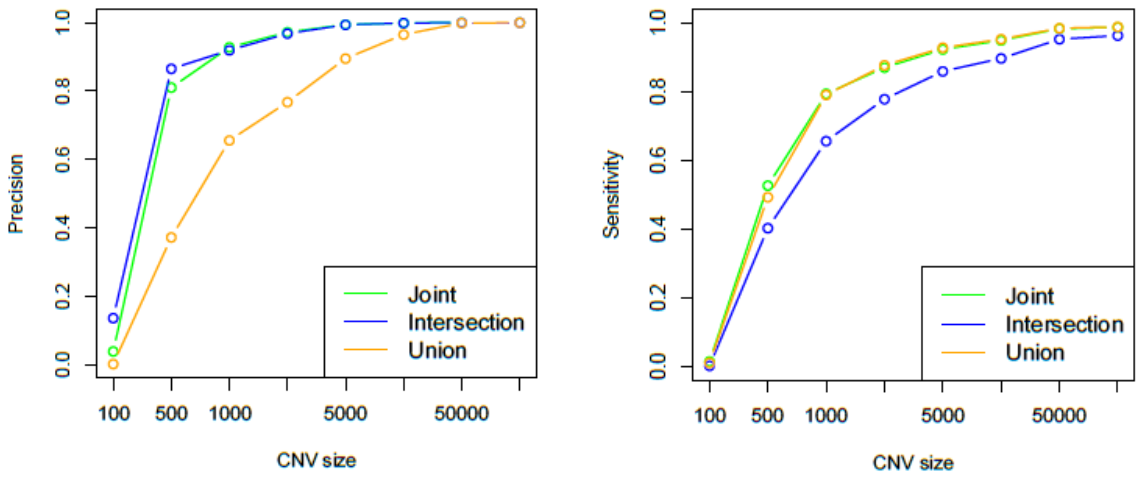


Figure 3.7 ***Results comparison between intersection or union and iCNV.*** *Precision and sensitivity analysis by in silico spike-in, comparing joint and intersection or union of two individual call set. Results show that joint calling has precision close to intersection and sensitivity close to union.*

BIBLIOGRAPHY

1. Gamazon ER, Stranger BE: **The impact of human copy number variation on gene expression.** *Briefings in Functional Genomics* 2015, **14:**352-357.
2. Hanks S, Coleman K, Reid S, Plaja A, Firth H, Fitzpatrick D, Kidd A, Mehes K, Nash R, Robin N, et al: **Constitutional aneuploidy and cancer predisposition caused by biallelic mutations in BUB1B.** *Nat Genet* 2004, **36:**1159-1161.
3. Vicente-Duenas C, Hauer J, Cobaleda C, Borkhardt A, Sanchez-Garcia I: **Epigenetic Priming in Cancer Initiation.** *Trends Cancer* 2018, **4:**408-417.
4. Burrell RA, McGranahan N, Bartek J, Swanton C: **The causes and consequences of genetic heterogeneity in cancer evolution.** *Nature* 2013, **501:**338-345.
5. Jiang Y, Qiu Y, Minn AJ, Zhang NR: **Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing.** *Proc Natl Acad Sci U S A* 2016, **113:**E5528-5537.
6. Deshwar AG, Vembu S, Yung CK, Jang GH, Stein L, Morris Q: **PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors.** *Genome Biol* 2015, **16:**35.
7. Zare H, Wang J, Hu A, Weber K, Smith J, Nickerson D, Song C, Witten D, Blau CA, Noble WS: **Inferring clonal composition from multiple sections of a breast cancer.** *PLoS Comput Biol* 2014, **10:**e1003703.
8. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, Laird PW, Onofrio RC, Winckler W, Weir BA, et al: **Absolute quantification of somatic DNA alterations in human cancer.** *Nat Biotechnol* 2012, **30:**413-421.
9. Li B, Li JZ: **A general framework for analyzing tumor subclonality using SNP array and DNA sequencing data.** *Genome Biol* 2014, **15:**473.
10. Oesper L, Mahmoody A, Raphael BJ: **THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data.** *Genome Biol* 2013, **14:**R80.
11. Ha G, Roth A, Khattra J, Ho J, Yap D, Prentice LM, Melnyk N, McPherson A, Bashashati A, Laks E, et al: **TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data.** *Genome Res* 2014, **24:**1881-1893.
12. Miller CA, White BS, Dees ND, Griffith M, Welch JS, Griffith OL, Vij R, Tomasson MH, Graubert TA, Walter MJ, et al: **SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution.** *PLoS Comput Biol* 2014, **10:**e1003665.
13. Navin NE: **The first five years of single-cell cancer genomics and beyond.** *Genome Res* 2015, **25:**1499-1507.
14. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, Cook K, Stepansky A, Levy D, Esposito D, et al: **Tumour evolution inferred by single-cell sequencing.** *Nature* 2011, **472:**90-94.
15. Wang Y, Waters J, Leung ML, Unruh A, Roh W, Shi X, Chen K, Scheet P, Vattathil S, Liang H, et al: **Clonal evolution in breast cancer revealed by single nucleus genome sequencing.** *Nature* 2014, **512:**155-160.
16. Gao R, Davis A, McDonald TO, Sei E, Shi X, Wang Y, Tsai PC, Casasent A, Waters J, Zhang H, et al: **Punctuated copy number evolution and clonal stasis in triple-negative breast cancer.** *Nat Genet* 2016, **48:**1119-1130.
17. Picelli S, Bjorklund AK, Faridani OR, Sagasser S, Winberg G, Sandberg R: **Smart-seq2 for sensitive full-length transcriptome profiling in single cells.** *Nat Methods* 2013, **10:**1096-1098.

18. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner MW: **Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells.** *Cell* 2015, **161:**1187-1201.
19. Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, Cahill DP, Nahed BV, Curry WT, Martuza RL, et al: **Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma.** *Science* 2014, **344:**1396-1401.
20. Chung W, Eum HH, Lee HO, Lee KM, Lee HB, Kim KT, Ryu HS, Kim S, Lee JE, Park YH, et al: **Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer.** *Nat Commun* 2017, **8:**15081.
21. Kim KT, Lee HW, Lee HO, Song HJ, Jeong da E, Shin S, Kim H, Shin Y, Nam DH, Jeong BC, et al: **Application of single-cell RNA sequencing in optimizing a combinatorial therapeutic strategy in metastatic renal cell carcinoma.** *Genome Biol* 2016, **17:**80.
22. Tirosh I, Izar B, Prakadan SM, Wadsworth MH, 2nd, Treacy D, Trombetta JJ, Rotem A, Rodman C, Lian C, Murphy G, et al: **Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq.** *Science* 2016, **352:**189-196.
23. Jerby-Arnon L, Shah P, Cuoco MS, Rodman C, Su MJ, Melms JC, Leeson R, Kanodia A, Mei S, Lin JR, et al: **A Cancer Cell Program Promotes T Cell Exclusion and Resistance to Checkpoint Blockade.** *Cell* 2018, **175:**984-997 e924.
24. Tirosh I, Venteicher AS, Hebert C, Escalante LE, Patel AP, Yizhak K, Fisher JM, Rodman C, Mount C, Filbin MG, et al: **Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma.** *Nature* 2016, **539:**309-313.
25. Venteicher AS, Tirosh I, Hebert C, Yizhak K, Neftel C, Filbin MG, Hovestadt V, Escalante LE, Shaw ML, Rodman C, et al: **Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq.** *Science* 2017, **355**.
26. Li H, Courtois ET, Sengupta D, Tan Y, Chen KH, Goh JJL, Kong SL, Chua C, Hon LK, Tan WS, et al: **Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors.** *Nat Genet* 2017, **49:**708-718.
27. Macaulay IC, Ponting CP, Voet T: **Single-Cell Multiomics: Multiple Measurements from Single Cells.** *Trends Genet* 2017, **33:**155-168.
28. Dey SS, Kester L, Spanjaard B, Bienko M, van Oudenaarden A: **Integrated genome and transcriptome sequencing of the same cell.** *Nat Biotechnol* 2015, **33:**285-289.
29. Macaulay IC, Haerty W, Kumar P, Li YI, Hu TX, Teng MJ, Goolam M, Saurat N, Coupland P, Shirley LM, et al: **G&T-seq: parallel sequencing of single-cell genomes and transcriptomes.** *Nat Methods* 2015, **12:**519-522.
30. Suva ML, Tirosh I: **Single-Cell RNA Sequencing in Cancer: Lessons Learned and Emerging Challenges.** *Mol Cell* 2019, **75:**7-12.
31. van Galen P, Hovestadt V, Wadsworth Ii MH, Hughes TK, Griffin GK, Battaglia S, Verga JA, Stephansky J, Pastika TJ, Lombardi Story J, et al: **Single-Cell RNA-Seq Reveals AML Hierarchies Relevant to Disease Progression and Immunity.** *Cell* 2019, **176:**1265-1281 e1224.
32. Nam AS, Kim KT, Chaligne R, Izzo F, Ang C, Taylor J, Myers RM, Abu-Zeinah G, Brand R, Omans ND, et al: **Somatic mutations and cell identity linked by Genotyping of Transcriptomes.** *Nature* 2019, **571:**355-360.
33. Raj A, Peskin CS, Tranchina D, Vargas DY, Tyagi S: **Stochastic mRNA synthesis in mammalian cells.** *PLoS Biol* 2006, **4:**e309.
34. Jiang Y, Zhang NR, Li M: **SCALE: modeling allele-specific gene expression by single-cell RNA sequencing.** *Genome Biol* 2017, **18:**74.
35. Padovan-Merhar O, Nair GP, Biaesch AG, Mayer A, Scarfone S, Foley SW, Wu AR, Churchman LS, Singh A, Raj A: **Single mammalian cells compensate for differences**

in cellular volume and DNA copy number through independent global transcriptional mechanisms. *Mol Cell* 2015, **58:**339-352.
36. Zafar H, Wang Y, Nakhleh L, Navin N, Chen K: **Monovar: single-nucleotide variant detection in single cells.** *Nat Methods* 2016, **13:**505-507.
37. Piskol R, Ramaswami G, Li JB: **Reliable identification of genomic variants from RNA-seq data.** *Am J Hum Genet* 2013, **93:**641-651.
38. Brennecke P, Anders S, Kim JK, Kolodziejczyk AA, Zhang X, Proserpio V, Baying B, Benes V, Teichmann SA, Marioni JC, Heisler MG: **Accounting for technical noise in single-cell RNA-seq experiments.** *Nat Methods* 2013, **10:**1093-1095.
39. Pierson E, Yau C: **ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis.** *Genome Biol* 2015, **16:**241.
40. Vallejos CA, Marioni JC, Richardson S: **BASiCS: Bayesian Analysis of Single-Cell Sequencing Data.** *PLoS Comput Biol* 2015, **11:**e1004333.
41. Ding B, Zheng L, Zhu Y, Li N, Jia H, Ai R, Wildberg A, Wang W: **Normalization and noise reduction for single cell RNA-seq experiments.** *Bioinformatics* 2015, **31:**2225-2227.
42. Qiu X, Hill A, Packer J, Lin D, Ma YA, Trapnell C: **Single-cell mRNA quantification and differential analysis with Census.** *Nat Methods* 2017, **14:**309-315.
43. Deng Q, Ramskold D, Reinius B, Sandberg R: **Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells.** *Science* 2014, **343:**193-196.
44. Eirew P, Steif A, Khattra J, Ha G, Yap D, Farahani H, Gelmon K, Chia S, Mar C, Wan A, et al: **Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution.** *Nature* 2015, **518:**422-426.
45. Gerlinger M, Rowan AJ, Horswell S, Math M, Larkin J, Endesfelder D, Gronroos E, Martinez P, Matthews N, Stewart A, et al: **Intratumor heterogeneity and branched evolution revealed by multiregion sequencing.** *N Engl J Med* 2012, **366:**883-892.
46. Shi YJ, Tsang JY, Ni YB, Tse GM: **Intratumoral Heterogeneity in Breast Cancer: A Comparison of Primary and Metastatic Breast Cancers.** *Oncologist* 2017, **22:**487-490.
47. Ribas A, Wolchok JD: **Cancer immunotherapy using checkpoint blockade.** *Science* 2018, **359:**1350-1355.
48. Schumacher TN, Schreiber RD: **Neoantigens in cancer immunotherapy.** *Science* 2015, **348:**69-74.
49. Rizvi NA, Hellmann MD, Snyder A, Kvistborg P, Makarov V, Havel JJ, Lee W, Yuan J, Wong P, Ho TS, et al: **Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer.** *Science* 2015, **348:**124-128.
50. Tumeh PC, Harview CL, Yearley JH, Shintaku IP, Taylor EJ, Robert L, Chmielowski B, Spasic M, Henry G, Ciobanu V, et al: **PD-1 blockade induces responses by inhibiting adaptive immune resistance.** *Nature* 2014, **515:**568-571.
51. Twyman-Saint Victor C, Rech AJ, Maity A, Rengan R, Pauken KE, Stelekati E, Benci JL, Xu B, Dada H, Odorizzi PM, et al: **Radiation and dual checkpoint blockade activate non-redundant immune mechanisms in cancer.** *Nature* 2015, **520:**373-377.
52. Benci JL, Johnson LR, Choa R, Xu Y, Qiu J, Zhou Z, Xu B, Ye D, Nathanson KL, June CH, et al: **Opposing Functions of Interferon Coordinate Adaptive and Innate Immune Responses to Cancer Immune Checkpoint Blockade.** *Cell* 2019, **178:**933-948 e914.
53. Patel SA, Minn AJ: **Combination Cancer Therapy with Immune Checkpoint Blockade: Mechanisms and Strategies.** *Immunity* 2018, **48:**417-433.
54. Goodman AM, Kato S, Bazhenova L, Patel SP, Frampton GM, Miller V, Stephens PJ, Daniels GA, Kurzrock R: **Tumor Mutational Burden as an Independent Predictor of**

Response to Immunotherapy in Diverse Cancers. *Mol Cancer Ther* 2017, **16:**2598-2608.
55. Rosenthal R, Cadieux EL, Salgado R, Bakir MA, Moore DA, Hiley CT, Lund T, Tanic M, Reading JL, Joshi K, et al: **Neoantigen-directed immune escape in lung cancer evolution.** *Nature* 2019, **567:**479-485.
56. Navin NE: **Cancer genomics: one cell at a time.** *Genome Biol* 2014, **15:**452.
57. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci U S A* 2005, **102:**15545-15550.
58. Naxerova K, Reiter JG, Brachtel E, Lennerz JK, van de Wetering M, Rowan A, Cai T, Clevers H, Swanton C, Nowak MA, et al: **Origins of lymphatic and distant metastases in human colorectal cancer.** *Science* 2017, **357:**55-60.
59. Wong JS, Warren LE, Bellon JR: **Management of the Regional Lymph Nodes in Early-Stage Breast Cancer.** *Semin Radiat Oncol* 2016, **26:**37-44.
60. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z, et al: **Supervised risk predictor of breast cancer based on intrinsic subtypes.** *J Clin Oncol* 2009, **27:**1160-1167.
61. Zhang JY, Zhang F, Hong CQ, Giuliano AE, Cui XJ, Zhou GJ, Zhang GJ, Cui YK: **Critical protein GAPDH and its regulatory mechanisms in cancer cells.** *Cancer Biol Med* 2015, **12:**10-22.
62. Tarrado-Castellarnau M, Diaz-Moralli S, Polat IH, Sanz-Pamplona R, Alenda C, Moreno V, Castells A, Cascante M: **Glyceraldehyde-3-phosphate dehydrogenase is overexpressed in colorectal cancer onset.** *Translational Medicine Communications* 2017, **2:**6.
63. Mann M, Cortez V, Vadlamudi RK: **Epigenetics of estrogen receptor signaling: role in hormonal cancer progression and therapy.** *Cancers (Basel)* 2011, **3:**1691-1707.
64. Green KA, Carroll JS: **Oestrogen-receptor-mediated transcription and the influence of co-factors and chromatin state.** *Nat Rev Cancer* 2007, **7:**713-722.
65. Dreijerink KM, Mulder KW, Winkler GS, Hoppener JW, Lips CJ, Timmers HT: **Menin links estrogen receptor activation to histone H3K4 trimethylation.** *Cancer Res* 2006, **66:**4929-4935.
66. Kim H, Heo K, Kim JH, Kim K, Choi J, An W: **Requirement of histone methyltransferase SMYD3 for estrogen receptor-mediated transcription.** *J Biol Chem* 2009, **284:**19867-19877.
67. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Jr., Kinzler KW: **Cancer genome landscapes.** *Science* 2013, **339:**1546-1558.
68. Tokheim CJ, Papadopoulos N, Kinzler KW, Vogelstein B, Karchin R: **Evaluating the evaluation of cancer driver genes.** *Proc Natl Acad Sci U S A* 2016, **113:**14330-14335.
69. Zhang W, Bojorquez-Gomez A, Velez DO, Xu G, Sanchez KS, Shen JP, Chen K, Licon K, Melton C, Olson KM, et al: **A global transcriptional network connecting noncoding mutations to changes in tumor gene expression.** *Nat Genet* 2018, **50:**613-620.
70. Cuykendall TN, Rubin MA, Khurana E: **Non-coding genetic variation in cancer.** *Curr Opin Syst Biol* 2017, **1:**9-15.
71. Singh M, Al-Eryani G, Carswell S, Ferguson JM, Blackburn J, Barton K, Roden D, Luciani F, Giang Phan T, Junankar S, et al: **High-throughput targeted long-read single cell sequencing reveals the clonal and transcriptional landscape of lymphocytes.** *Nat Commun* 2019, **10:**3120.
72. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: **STAR: ultrafast universal RNA-seq aligner.** *Bioinformatics* 2013, **29:**15-21.

73. Li B, Dewey CN: **RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.** *BMC Bioinformatics* 2011, **12:**323.
74. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.** *Genome Res* 2010, **20:**1297-1303.
75. Skelly DA, Johansson M, Madeoy J, Wakefield J, Akey JM: **A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data.** *Genome Res* 2011, **21:**1728-1737.
76. Ward JH: **Hierarchical Grouping to Optimize an Objective Function.** *Journal of the American Statistical Association* 1963, **58:**236-&.
77. Goutte C, Hansen LK, Liptrot MG, Rostrup E: **Feature-space clustering for fMRI meta-analysis.** *Hum Brain Mapp* 2001, **13:**165-183.
78. Urrutia E, Chen H, Zhou Z, Zhang NR, Jiang Y: **Integrative pipeline for profiling DNA copy number and inferring tumor phylogeny.** *Bioinformatics* 2018, **34:**2126-2128.
79. Pfeiffer F, Grober C, Blank M, Handler K, Beyer M, Schultze JL, Mayer G: **Systematic evaluation of error rates and causes in short samples in next-generation sequencing.** *Sci Rep* 2018, **8:**10950.
80. Li B, Chen W, Zhan X, Busonero F, Sanna S, Sidore C, Cucca F, Kang HM, Abecasis GR: **A likelihood-based framework for variant calling and de novo mutation detection in families.** *PLoS Genet* 2012, **8:**e1002944.
81. Schliep KP: **phangorn: phylogenetic analysis in R.** *Bioinformatics* 2011, **27:**592-593.
82. Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, Slichter CK, Miller HW, McElrath MJ, Prlic M, et al: **MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data.** *Genome Biol* 2015, **16:**278.
83. Korthauer KD, Chu LF, Newton MA, Li Y, Thomson J, Stewart R, Kendziorski C: **A statistical approach for identifying differential distributions in single-cell RNA-seq experiments.** *Genome Biol* 2016, **17:**222.
84. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R: **Integrating single-cell transcriptomic data across different conditions, technologies, and species.** *Nat Biotechnol* 2018, **36:**411-420.
85. Wang K, Li M, Hakonarson H: **ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data.** *Nucleic Acids Res* 2010, **38:**e164.
86. Karosiene E, Lundegaard C, Lund O, Nielsen M: **NetMHCcons: a consensus method for the major histocompatibility complex class I predictions.** *Immunogenetics* 2012, **64:**177-186.
87. Stuart T, Satija R: **Integrative single-cell analysis.** *Nat Rev Genet* 2019, **20:**257-272.
88. Peterson VM, Zhang KX, Kumar N, Wong J, Li L, Wilson DC, Moore R, McClanahan TK, Sadekova S, Klappenbach JA: **Multiplexed quantification of proteins and transcripts in single cells.** *Nat Biotechnol* 2017, **35:**936-939.
89. Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, Satija R, Smibert P: **Simultaneous epitope and transcriptome measurement in single cells.** *Nat Methods* 2017, **14:**865-868.
90. Wang X, Allen WE, Wright MA, Sylwestrak EL, Samusik N, Vesuna S, Evans K, Liu C, Ramakrishnan C, Liu J, et al: **Three-dimensional intact-tissue sequencing of single-cell transcriptional states.** *Science* 2018, **361**.
91. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, Bodenmiller B, Campbell P, Carninci P, Clatworthy M, et al: **The Human Cell Atlas.** *Elife* 2017, **6**.
92. Villani AC, Satija R, Reynolds G, Sarkizova S, Shekhar K, Fletcher J, Griesbeck M, Butler A, Zheng S, Lazo S, et al: **Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors.** *Science* 2017, **356**.

93. Liu Y, Beyer A, Aebersold R: **On the Dependency of Cellular Protein Levels on mRNA Abundance.** *Cell* 2016, **165:**535-550.
94. Svensson V, Natarajan KN, Ly LH, Miragaia RJ, Labalette C, Macaulay IC, Cvejic A, Teichmann SA: **Power analysis of single-cell RNA-sequencing experiments.** *Nat Methods* 2017, **14:**381-387.
95. Zhao BS, Roundtree IA, He C: **Post-transcriptional gene regulation by mRNA modifications.** *Nat Rev Mol Cell Biol* 2017, **18:**31-42.
96. Jackson RJ, Hellen CU, Pestova TV: **The mechanism of eukaryotic translation initiation and principles of its regulation.** *Nat Rev Mol Cell Biol* 2010, **11:**113-127.
97. Mowen KA, David M: **Unconventional post-translational modifications in immunological signaling.** *Nat Immunol* 2014, **15:**512-520.
98. Schwartz AL: **Cell biology of intracellular protein trafficking.** *Annu Rev Immunol* 1990, **8:**195-229.
99. Roux PP, Topisirovic I: **Signaling Pathways Involved in the Regulation of mRNA Translation.** *Mol Cell Biol* 2018, **38**.
100. Wang J, Agarwal D, Huang M, Hu G, Zhou Z, Ye C, Zhang NR: **Data denoising with transfer learning in single-cell transcriptomics.** *Nat Methods* 2019, **16:**875-878.
101. Webb S: **Deep learning for biology.** *Nature* 2018, **554:**555-557.
102. Tang B, Pan Z, Yin K, Khateeb A: **Recent Advances of Deep Learning in Bioinformatics and Computational Biology.** *Front Genet* 2019, **10:**214.
103. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N: **Deep generative modeling for single-cell transcriptomics.** *Nat Methods* 2018, **15:**1053-1058.
104. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, 3rd, Hao Y, Stoeckius M, Smibert P, Satija R: **Comprehensive Integration of Single-Cell Data.** *Cell* 2019, **177:**1888-1902 e1821.
105. Martins PS, Brunialti MK, Martos LS, Machado FR, Assuncao MS, Blecher S, Salomao R: **Expression of cell surface receptors and oxidative metabolism modulation in the clinical continuum of sepsis.** *Crit Care* 2008, **12:**R25.
106. Chen L, Flies DB: **Molecular mechanisms of T cell co-stimulation and co-inhibition.** *Nat Rev Immunol* 2013, **13:**227-242.
107. Fromm P, Papadimitrious M, Hsu J, Larsen SR, Gibson J, Bradstock K, Kupresanin F, Clark G, Hart DNJ: **CD16+Dendritic Cells Are a Unique Myeloid Antigen Presenting Cell Population.** *Blood* 2016, **128**.
108. D'Arena G, Musto P, Cascavilla N, Di Giorgio G, Fusilli S, Zendoli F, Carotenuto M: **Flow cytometric characterization of human umbilical cord blood lymphocytes: immunophenotypic features.** *Haematologica* 1998, **83:**197-203.
109. Clavarino G, Delouche N, Vettier C, Laurin D, Pernollet M, Raskovalova T, Cesbron JY, Dumestre-Perard C, Jacob MC: **Novel Strategy for Phenotypic Characterization of Human B Lymphocytes from Precursors to Effector Cells by Flow Cytometry.** *Plos One* 2016, **11**.
110. Van Acker HH, Capsomidis A, Smits EL, Van Tendeloo VF: **CD56 in the Immune System: More Than a Marker for Cytotoxicity?** *Front Immunol* 2017, **8:**892.
111. Tsukerman P, Stern-Ginossar N, Yamin R, Ophir Y, Stanietsky AM, Mandelboim O: **Expansion of CD16 positive and negative human NK cells in response to tumor stimulation.** *Eur J Immunol* 2014, **44:**1517-1525.
112. Poli A, Michel T, Theresine M, Andres E, Hentges F, Zimmer J: **CD56(bright) natural killer (NK) cells: an important NK cell subset.** *Immunology* 2009, **126:**458-465.
113. Wendt K, Wilk E, Buyny S, Buer J, Schmidt RE, Jacobs R: **Gene and protein characteristics reflect functional diversity of CD56(dim) and CD56(bright) NK cells.** *Journal of Leukocyte Biology* 2006, **80:**1529-1541.

114. d'Angeac AD, Monier S, Pilling D, Travaglio-Encinoza A, Reme T, Salmon M: **CD57+ T lymphocytes are derived from CD57- precursors by differentiation occurring in late immune responses.** *Eur J Immunol* 1994, **24:**1503-1511.
115. Musha N, Yoshida Y, Sugahara S, Yamagiwa S, Koya T, Watanabe H, Hatakeyama K, Abo T: **Expansion of CD56+ NK T and gamma delta T cells from cord blood of human neonates.** *Clin Exp Immunol* 1998, **113:**220-228.
116. Dalle JH, Menezes J, Wagner E, Blagdon M, Champagne J, Champagne MA, Duval M: **Characterization of cord blood natural killer cells: implications for transplantation and neonatal infections.** *Pediatr Res* 2005, **57:**649-655.
117. Pollyea DA, Jordan CT: **Therapeutic targeting of acute myeloid leukemia stem cells.** *Blood* 2017, **129:**1627-1635.
118. McKenzie MD, Ghisi M, Oxley EP, Ngo S, Cimmino L, Esnault C, Liu RJ, Salmon JM, Bell CC, Ahmed N, et al: **Interconversion between Tumorigenic and Differentiated States in Acute Myeloid Leukemia.** *Cell Stem Cell* 2019, **25:**258-+.
119. Geissmann F, Manz MG, Jung S, Sieweke MH, Merad M, Ley K: **Development of Monocytes, Macrophages, and Dendritic Cells.** *Science* 2010, **327:**656-661.
120. Jang JH, Yoo EH, Kim HJ, Kim DH, Jung CW, Kim SH: **Acute myeloid leukemia with del(X)(p21) and cryptic RUNX1/RUNX1T1 from ins(8;21)(q22;q22q22) revealed by atypical FISH signals.** *Ann Clin Lab Sci* 2010, **40:**80-84.
121. Moroi K, Sato T: **Comparison between procaine and isocarboxazid metabolism in vitro by a liver microsomal amidase-esterase.** *Biochem Pharmacol* 1975, **24:**1517-1521.
122. Shang L, Chen X, Liu Y, Cai X, Shi Y, Shi L, Li Y, Song Z, Zheng B, Sun W, et al: **The immunophenotypic characteristics and flow cytometric scoring system of acute myeloid leukemia with t(8;21) (q22;q22); RUNX1-RUNX1T1.** *Int J Lab Hematol* 2019, **41:**23-31.
123. Naik J, Themeli M, de Jong-Korlaar R, Ruiter RWJ, Poddighe PJ, Yuan HP, Bruijn JDD, Ossenkoppele GJ, Zweegman S, Smit L, et al: **CD38 as a therapeutic target for adult acute myeloid leukemia and T-cell acute lymphoblastic leukemia.** *Haematologica* 2019, **104:**E100-E103.
124. Eveillard M, Floc'h V, Robillard N, Debord C, Wuilleme S, Garand R, Rialland F, Thomas C, Peterlin P, Guillaume T, et al: **CD38 Expression in B-Lineage Acute Lymphoblastic Leukemia, a Possible Target for Immunotherapy.** *Blood* 2016, **128**.
125. An GZ: **The effects of adding noise during backpropagation training on a generalization performance.** *Neural Computation* 1996, **8:**643-674.
126. Reed R, MarksII RJ: *Neural smithing: supervised learning in feedforward artificial neural networks.* Mit Press; 1999.
127. Andrews TS, Hemberg M: **False signals induced by single-cell imputation.** *F1000Res* 2018, **7:**1740.
128. LeCun Y, Bengio Y, Hinton G: **Deep learning.** *Nature* 2015, **521:**436-444.
129. Kingma D, Ba J: **Adam: a method for stochastic optimization (2014).** *arXiv preprint arXiv:14126980* 2015, **15**.
130. Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, Aburatani H, Jones KW, Tyler-Smith C, Hurles ME, et al: **Copy number variation: new insights in genome diversity.** *Genome Res* 2006, **16:**949-961.
131. McCarroll SA, Altshuler DM: **Copy-number variation and association studies of human disease.** *Nat Genet* 2007, **39:**S37-42.
132. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, et al: **Global variation in copy number in the human genome.** *Nature* 2006, **444:**444-454.

133. Diskin SJ, Hou C, Glessner JT, Attiyeh EF, Laudenslager M, Bosse K, Cole K, Mosse YP, Wood A, Lynch JE, et al: **Copy number variation at 1q21.1 associated with neuroblastoma.** *Nature* 2009, **459:**987-991.
134. Glessner JT, Wang K, Cai G, Korvatska O, Kim CE, Wood S, Zhang H, Estes A, Brune CW, Bradfield JP, et al: **Autism genome-wide copy number variation reveals ubiquitin and neuronal genes.** *Nature* 2009, **459:**569-573.
135. McCarroll SA, Huett A, Kuballa P, Chilewski SD, Landry A, Goyette P, Zody MC, Hall JL, Brant SR, Cho JH, et al: **Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease.** *Nat Genet* 2008, **40:**1107-1112.
136. Fromer M, Moran JL, Chambert K, Banks E, Bergen SE, Ruderfer DM, Handsaker RE, McCarroll SA, O'Donovan MC, Owen MJ, et al: **Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth.** *Am J Hum Genet* 2012, **91:**597-607.
137. Jiang Y, Oldridge DA, Diskin SJ, Zhang NR: **CODEX: a normalization and copy number variation detection method for whole exome sequencing.** *Nucleic Acids Res* 2015, **43:**e39.
138. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, Hakonarson H, Bucan M: **PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data.** *Genome Res* 2007, **17:**1665-1674.
139. Abyzov A, Urban AE, Snyder M, Gerstein M: **CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing.** *Genome Res* 2011, **21:**974-984.
140. Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, et al: **High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays.** *Nat Genet* 1998, **20:**207-211.
141. Carter NP: **Methods and strategies for analyzing copy number variation using DNA microarrays.** *Nat Genet* 2007, **39:**S16-21.
142. Chiang DY, Getz G, Jaffe DB, O'Kelly MJ, Zhao X, Carter SL, Russ C, Nusbaum C, Meyerson M, Lander ES: **High-resolution mapping of copy-number alterations with massively parallel sequencing.** *Nat Methods* 2009, **6:**99-103.
143. Zhao M, Wang Q, Wang Q, Jia P, Zhao Z: **Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives.** *BMC Bioinformatics* 2013, **14 Suppl 11:**S1.
144. Zhou Z, Wang W, Wang LS, Zhang NR: **Integrative DNA copy number detection and genotyping from sequencing and array-based platforms.** *Bioinformatics* 2018, **34:**2349-2355.
145. Aguirre M, Rivas MA, Priest J: **Phenome-wide Burden of Copy-Number Variation in the UK Biobank.** *Am J Hum Genet* 2019, **105:**373-383.
146. Takahashi PY, Jenkins GD, Welkie BP, McDonnell SK, Evans JM, Cerhan JR, Olson JE, Thibodeau SN, Cicek MS, Ryu E: **Association of mitochondrial DNA copy number with self-rated health status.** *Appl Clin Genet* 2018, **11:**121-127.
147. Dean J, Ghemawat S: **Mapreduce: Simplified data processing on large clusters.** *Communications of the Acm* 2008, **51:**107-113.
148. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25:**2078-2079.
149. Packer JS, Maxwell EK, O'Dushlaine C, Lopez AE, Dewey FE, Chernomorsky R, Baras A, Overton JD, Habegger L, Reid JG: **CLAMMS: a scalable algorithm for calling common and rare copy number variants from exome sequencing data.** *Bioinformatics* 2016, **32:**133-135.

150. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, Wang D, Masys DR, Roden DM, Crawford DC: **PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations.** *Bioinformatics* 2010, **26:**1205-1210.